

For Reference

NOT TO BE TAKEN FROM THIS ROOM

Ex libris
UNIVERSITATIS
ALBERTAENSIS



THE UNIVERSITY OF ALBERTA

AN EVALUATION OF TWO STANDARDIZED TESTS WITH REGARD
TO THEIR SUITABILITY FOR MEASURING ACHIEVEMENT IN
TWO MATHEMATICS PROGRAMS

by



ROBERT B. DARGAVEL

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE
OF MASTER OF EDUCATION

DEPARTMENT OF ELEMENTARY EDUCATION

EDMONTON, ALBERTA

FALL, 1971

THE UNIVERSITY OF ALBERTA
FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommended to the Faculty of Graduate Studies for acceptance, a thesis entitled "An Evaluation of Two Standardized Tests With Regard to Their Suitability for Measuring Achievement in Two Mathematics Programs" submitted by Robert B. Dargavel in partial fulfilment of the requirements for the degree of Master of Education.

ABSTRACT

This study attempted to determine the suitability of two recently published standardized tests, the Cooperative Primary Test, Mathematics Form 12A and the Modern Mathematics Understanding Test, Primary Form C (the SRA Test), for measuring achievement in mathematics in two modern mathematics programs currently in use in the Province of Alberta.

One hundred and twenty pupils participated in the study. These pupils were selected from six schools, three of which followed the mathematics program of the Addison-Wesley (AW) Text Series and three which followed the Seeing Through Arithmetic Text Series (STA). Thirty pupils were chosen from each of grades one and two of the two mathematics programs. The pupils were chosen by the school personnel on the basis of mathematics ability. They were asked to choose equal numbers of pupils in each of three ability groups at each grade level.

The content validity of the two tests with respect to the two grade levels of each of the mathematics programs was assessed by submitting the tests to the scrutiny of a committee of judges. The content validity coefficients so obtained were compared by treating them as independent proportions.

The two tests were then administered to each of the one hundred and twenty pupils by the researcher. The achievement scores so obtained were then compared using one-way analysis of variance and one and two-way analysis of covariance. Scores of pupils in the AW program were compared with scores of pupils in the STA program

at each grade level.

These results were then interpreted in terms of the content validity, difficulty and number of valid items of each test for the two programs being compared.

The SRA Test is machine markable. A comparison was made of the scores of pupils answering the test using this format with those answering in handscoring format.

The number of unsuitable items on each test for each of the four mathematics programs was identified and compared. These results were then interpreted in terms of item difficulty to aid in the determination of the suitability of these tests for use in conjunction with these mathematics programs.

On the basis of the evidence gathered in this study the following conclusions may be made concerning the tests being evaluated in this study. The Cooperative Test could be used as a post-test in grade one and as a pre-test in grade two. It would not be suitable as a post-test in grade two. The SRA Test could be used as a post-test in grade two and a pre-test in grade three. It is not suitable for use in grade one. All decisions concerning the use of these tests would have to be based on a value judgment by prospective users, concerning the suitability of the content validity of the tests for the intended purposes of testing.

ACKNOWLEDGMENTS

The writer wishes to express grateful appreciation to Dr. K. Allen Neufeld, for his invaluable guidance and assistance both prior to and during the preparation of this thesis. Appreciation is also extended to the members of the Thesis Committee, Professor N. M. Purvis, and Dr. T. E. Kieren for their suggestions.

The writer is indebted to Dr. George Cathcart for his advice and assistance in setting up the statistical analysis used in this study and to the persons who acted as judges in the assessment of content validity.

Grateful acknowledgement is expressed to the Edmonton Separate School Board without whose cooperation this study would have been impossible.

Finally, this writer wishes to express his deep appreciation to Miss Helen Milton, for her patience and encouragement during the past year.

TABLE OF CONTENTS

Chapter		Page
1.	INTRODUCTION TO THE STUDY.....	1
	General Introduction.....	1
	The Purpose of the Study.....	2
	Variables and Instruments.....	2
	Definition of Terms.....	3
	Hypotheses.....	5
	Significance of the Study.....	6
	Limitations of the Study.....	7
	The Statistical Design.....	7
	Outline of the Report.....	8
2.	RELATED LITERATURE.....	9
	Introduction.....	9
	Justification of the Meaning Method.....	9
	The Importance of Measuring Meanings.....	15
	Attempts to Evaluate Meaning.....	19
	Standardized Tests: Uses and Some Criteria for Selection.....	26
	Summary.....	26
3.	METHODS AND PROCEDURE.....	27
	The Sample.....	27
	The Instruments.....	28
	The Cooperative Primary Test, Mathematics, Form 12A	28
	The Modern Mathematics Understanding Test, Primary Form C (The SRA Test).....	35

Chapter	Page
Procedure	37
Statistical Procedures.....	41
Intelligence as a Covariate.....	43
4. THE FINDINGS OF THE STUDY.....	46
The Findings With Respect to Content Validity.....	46
Results With Respect to the Hypotheses.....	57
Results of Further Analysis.....	75
Summary of the Results.....	80
5. SUMMARY, CONCLUSIONS, IMPLICATIONS AND SUGGESTIONS FOR FURTHER RESEARCH.....	85
Summary.....	85
Conclusions.....	88
Conclusions With Respect to the Major Hypotheses...	88
Conclusions With Respect to Additional Analysis....	95
Implications.....	95
Suggestions for Further Research.....	97
BIBLIOGRAPHY.....	99
APPENDIX A List of Judges.....	104
APPENDIX B Objectives of the Mathematics Programs Used in This Study.....	106
APPENDIX C Judges Matching of Items to Objectives in the Assessment of Content Validity.....	121
APPENDIX D Tables Showing Unsuitable Items and their Difficulty Indexes on the Basic Facts and Number Numeration Subtests.....	129

Chapter		Page
APPENDIX E	Tables Showing the Items and their Difficulty Indexes.....	142
APPENDIX F	The Cooperative 23A Test.....	151

LIST OF TABLES

Table	Page
1. Analysis of Variance Comparing by Grade Level Mean I.Q. Scores of the AW and STA Groups.....	29
2. Analysis of Variance Comparing by Grade Level the Three Ability Groups of the AW Program to the Three Ability Groups of the STA Program	30
3. The Number of Items on Each of the Various Subtests Identified by the Researcher of (a) The Cooperative Test (b) The SRA Test.....	38
4. Numerical Description of the Instruments Used In This Study.....	39
5. Correlation Between I.Q. and Achievement Score for the Pupils Used in the Study.....	45
6. Results of the Content Validity Assessment	
(a) Cooperative Test and the AW1 Program.....	49
(b) Cooperative Test and the STA1 Program.....	50
(c) Cooperative Test and the AW2 Program.....	51
(d) Cooperative Test and the STA2 Program.....	52
(e) SRA Test and the AW1 Program.....	53
(f) SRA Test and the STA1 Program.....	54
(g) SRA Test and the AW2 Program.....	55
(h) SRA Test and the STA2 Program.....	56

Table		Page
7.	Comparison of the Content Validity of the Cooperative Test and the SRA Test With Reference to the Objectives of (a) the AW1 Program (b) the STA1 Program (c) the AW2 Program (d) the STA2 Program.....	59
8.	Analysis of Covariance on the Criterion of Mathematics Achievement of Pupils in the Two Grade of the AW Program and the STA Program with I.Q. as Covariate.....	61
9.	Comparison of the Mean Achievement Scores Attained on the Cooperative Test and the SRA Test by pupils in (a) the AW1 Program (b) the STA1 Program (c) the AW2 Program (d) the STA2 Program.....	63
10.	Analysis of Covariance with I.Q. as the Covariate on the Criterion of Mathematics Achievement for the High, Average and Low Ability Groups (Grade One).....	65
11.	Analysis of Covariance, with I.Q. as the Covariate, on the Criterion of Mathematics Achievement for the High, Average and Low Ability Groups (Grade Two).....	67
12.	Comparison of Results Attained by Grade One Pupils in the STA Program and the AW Program on Two Subtests of of (a) the Cooperative Test (b) the SRA Test.....	69
13.	Comparison of Results attained by Grade Two Pupils in the STA Program and the AW Program on Two Subtests of (a) the Cooperative Test (b) the SRA Test.....	71

Table		Page
14.	Results of the Analysis of Covariance, Using IQ as Covariate, Comparing the Pupils Answering the SRA Test in Machine-Markable Format and Pupils Answering in Hand Scoring Format at Both Grade Levels on the Criterion of Mathematics Achievement....	72
15.	Comparison of the number of items judged "Not Suitable for Pupils in the (a) AW1 Program (b) the STA1 Program (c) the AW2 Program (d) the STA2 Program.....	74
16.	Summary of Comparisons Made Between the Cooperative and SRA Tests:- (a) With Respect to the AW1 Program..... (b) With Respect to the STA1 Program..... (c) With Respect to the AW2 Program..... (d) With Respect to the STA2 Program.....	76 77 78 79
17.	Summary of Information Concerning the Cooperative Test.....	84
18.	Summary of Information Concerning the SRA Test.....	85

Chapter 1

INTRODUCTION TO THE STUDY

GENERAL INTRODUCTION

Evaluation is an essential part of any educational program. Johnson (1961) calls evaluation the quality control of the educational program. It is the means by which the quality of our programs can be constantly improved. If evaluation is to be meaningful the instrument chosen must accurately measure the objectives of the program.

In view of the great change in emphasis in the mathematics curriculum, from computational skill to the meanings in mathematics, this choice of an appropriate instrument to assess the program has become exceedingly difficult. As Romberg (1966) states

One of the major problems in the now decade old mathematics curriculum revolution has been the lack of appropriate assessment devices for measuring the effectiveness of modern mathematics programs. (p. 349)

The magnitude of this problem was brought more sharply into focus by two recent occurrences.

The first of these is in connection with the curriculum experiment in Individually Prescribed Instruction (I.P.I.) being carried out in Alberta by the Alberta Human Resources Research Council. In September, 1971, the experiment will begin its third and final year and at this point in time they do not have an adequate instrument for measuring achievement at the grade one and grade two levels.

Similarly the Department of Education of the Province of Alberta is searching for a mathematics test which will measure

achievement. Three textbook series presenting modern mathematics programs have been authorized as primary references based on the provincial curricular guidelines for mathematics. Instruments which were adequate for curricula of the past are no longer appropriate. Standardized tests, if suitable, can be used very effectively in both of the above situations. The question arises then, are there suitable standardized tests available that will effectively measure achievement in modern mathematics programs?

THE PURPOSE OF THE STUDY

This study attempts to evaluate two recently developed standardized tests that are available for measuring mathematics achievement at the primary level. This evaluation will give some indication of the adequacy of these tests for use in conjunction with the existing mathematics programs in the Province of Alberta.

More specifically the purpose of the study was (1) to determine the content validity of the two tests with respect to two recognized text series used in the primary grades in the Province of Alberta, (2) to compare the results of pupils using the two text series on each test, (3) to determine problems in administering the tests and difficulties the pupils may have in interpreting the instructions given to them during the test, (4) to ascertain primary children's ability to handle the machine-marking format available on one of the tests.

VARIABLES AND INSTRUMENTS

Independent Variables

1. Ability Groupings

The schools were asked to choose an equal number of pupils

in each of the following groups (1) high achievers in mathematics (2) average achievers in mathematics (3) low achievers in mathematics. The schools were further asked not to include any exceptional children.

2. Mathematics Programs

The two mathematics programs to be used in this study are those which follow, firstly, the Seeing Through Arithmetic Series (STA1 and STA2) published by the W. L. Gage Co. Ltd., Toronto and, secondly, the Elementary School Mathematics Series, revised edition, books one and two (AW1 and AW2) published by the Addison-Wesley Co. Ltd., Don Mills, Ontario.

3. Grade Level

The two levels at which achievement will be measured are grades one and two.

Dependent Variables

1. Mathematics Achievement

Achievement scores will be obtained by using two standardized tests, the Cooperative Primary Test, Mathematics, Form 12A published by Educational Testing Service, Berkely, California, and the SRA Modern Mathematics Understanding Test, Primary Level, Form C published by Science Research Associates (Canada) Limited, Don Mills, Ontario.

DEFINITION OF TERMS

(1) Content validity of a test is defined to be the percentage of the objectives of a particular mathematics program tested by a particular test.

(2) Item difficulty is defined mathematically as follows:

$$\text{Dif} = \frac{N_r}{N_t - N_f}$$

where N_r is the number of students correctly answering an item
 N_t is the total number of students writing the test

N_f is the number of students not responding to the test item because they did not finish the examination.

The difficulty index has a range of $0 \leq \text{Dif} \leq 1$. A difficulty index of 0 indicates that no correct responses were given for an item. An index of 1.0 indicates no incorrect responses.

(3) The Biserial Correlation of a test item is defined mathematically as

$$\text{Biserial Correl} = \frac{M_r - M_w}{S_x^2} \times \frac{P}{Z}$$

where M_r is the mean of the students answering the item correctly
 M_w is the mean of the students answering the item incorrectly
 S_x^2 is the total test variance
 P is the item difficulty index
 Z is the ordinate in the unit normal distribution corresponding to the proportion p .

(4) Item Reliability Index is defined mathematically as

$$\text{Item Rel. Index} = \text{Biserial Correl} \times \frac{1}{\sqrt{1 - \text{Diff}}}$$

This index has a range of $-0.5 \leq \text{Index} \leq 0.5$

(5) Item suitability: An item will be judged to be suitable if it has an Item Reliability Index greater than or equal to .1

(6) Valid Item: An item will be called valid if the material which it tests is material taught to the students writing the tests. That is, it tests material contained in the mathematics program taken by the students writing the test.

(7) Meaningful Approach (the meaning method): The "meaningful approach" is a method of teaching which offers learners a planned, sequential systematic series of experiences in which they have opportunities to develop understanding of the basic principles of the subject being taught.

(8) Understandings: Understandings are notions resulting from the comprehension of relationships between numbers and between operations on numbers. The term suggests an awareness of patterns in mathematics and the ability to represent numbers with symbols. In this study, "understandings" is not used to refer to computational abilities.

HYPOTHESES

(1) There is no significant difference between the content validity of the Cooperative Test and SRA Test with reference to the objectives of (a) the STA1 program (b) the AW1 program (c) the STA2 program (d) the AW2 program.

(2) There is no significant difference between the mean achievement scores of pupils in the STA1 program and pupils in the AW1 program as measured by (a) the Cooperative Test (b) the SRA Test.

(3) There is no significant difference between the mean achievement scores of pupils in the STA 2 program and pupils in the AW2 program as measured by (a) the Cooperative Test (b) the SRA Test.

(4) In measuring achievement, there is no significant difference between the mean score obtained on the Cooperative Test and the mean score obtained on the SRA Test for pupils in (a) the STA1 program (b) the AW1 program (c) the STA2 program (d) the AW2 program.

(5) There is no significant difference between the mean achievement scores of pupils in (a) the high ability groups (b) the middle ability groups (c) the low ability groups of the STA 1 program and the AW1 program as measured by both the Cooperative Test and the SRA Test.

(6) There is no significant difference between the mean achievement scores of pupils in (a) the high ability groups (b) the middle

ability groups (c) the low ability groups of the STA2 program and the AW2 program as measured by both the Cooperative Test and the SRA Test.

(7) There is no significant difference between the mean achievement scores of pupils in the STA1 program and the AW1 program on the basic facts subtest of (a) the Cooperative Test (b) the SRA Test.

(8) There is no significant difference between the mean achievement scores of pupils in the STA2 program and the AW2 program on the basic facts subtest of (a) the Cooperative Test (b) the SRA Test.

(9) There is no significant difference between the mean achievement scores of pupils in the STA1 program and AW1 program on the number-numeration subtest of (a) the Cooperative Test (b) the SRA Test.

(10) There is no significant difference between the mean achievement scores of pupils in the STA2 program and AW2 program on the number-numeration subtest of (a) the Cooperative Test (b) the SRA Test.

(11) There is no significant difference between the mean achievement scores of pupils writing the machine markable test and those writing the hand scored test in (a) grade one (b) grade two.

(12) There is no difference in the percentage of items judged suitable on the Cooperative Test compared to the percentage of items judged suitable on the SRA Test at each grade level.

SIGNIFICANCE OF THE STUDY

The mathematics curriculum within our schools is constantly being evaluated by both local boards and provincial departments of

education. Part of this evaluation is the general assessment of pupil achievement of the instructional objectives set down in the curriculum guides for each division within the school structure. Suitable standardized tests could be very useful in such an assessment.

It is hoped that the results of this study will possibly be of use in the selection of a standardized test for use in measuring mathematics achievement in the primary grades.

Further it may be that these results would be of assistance in the construction of an achievement test by indicating types of items that are suitable at these grade levels, and also those that are unsuitable. Areas in the curriculum not adequately tested by existing instruments could also be indicated.

LIMITATIONS OF THE STUDY

(1) Since content validity may vary significantly from classroom to classroom the results of this section of the study may not be generalized.

(2) Since the sample is small, it may not be representative of the total population in any one program.

(3) The pupils were chosen by the schools. Thus the assumption that the ability groups being compared are equivalent may not be valid.

THE STATISTICAL DESIGN

The design is reported in detail in Chapter III. A brief overview is given here for the purpose of orientation.

The content validity of the two standardized tests selected to be evaluated was assessed by having a committee of four judges scrutinize the tests. On the basis of this scrutiny the percentage of

objectives being tested in each of the four programs was determined.

A sample of one hundred and twenty children was selected from the Edmonton Separate School System, sixty in grade one and sixty in grade two. The sixty children in each grade were divided evenly between the two mathematics programs. They were also chosen so that there would be equal numbers in each achievement level in each grade. Each test to be evaluated was administered by the researcher during the month of April, 1971. The results were analyzed by (a) comparing the results of children in a particular program on the two tests, (b) comparing the results of children in different programs on a particular test (c) comparing the results of children in each ability group in both programs on each test (d) comparing machine-markable results with hand-scored results.

An item analysis was used to determine the number of unsuitable items on each test at each grade level. The percentages of unsuitable items were then compared.

OUTLINE OF THE REPORT

Having introduced the problem and the study in this chapter, a review of the literature related to the problem will be presented in Chapter II. Chapter III will include a detailed description of the experimental design and a discussion of the statistical procedures used in analyzing the data. The results of the data analysis will be reported in Chapter IV. Finally, a summary of the study including the conclusions, implications and suggestions for further research will be presented in Chapter V.

Chapter 2

RELATED LITERATURE

INTRODUCTION

This chapter is divided into four sections. The first is a justification of the teaching mathematics for understanding rather than merely for computational skill; the second presents a rationale for testing children's understandings of meanings in mathematics; the third section presents attempts that have been made to test understandings in mathematics; the fourth section deals with standardized achievement tests, their use and the criteria for choosing such a test.

JUSTIFICATION OF THE MEANING METHOD

This section is also divided into four parts. The first is a brief history of the teaching of mathematics; the second part deals with some results from psychology dealing with teaching for understanding; the third deals with experiments carried out in the area of mathematics education dealing with the meaning method; the fourth section presents some views of mathematics educators on the meaning method. The purpose of this section will be to show that the modern mathematics reform movement is not a fad or temporary preoccupation with novel approaches to mathematics. This movement validly reflects, as have other major changes in educational programs before it, the continued need to change educational programs to keep them in line with the developing state of knowledge in the field, as well as with the needs of society.

In the period preceding 1890, the central aim of education

was to improve man himself. The common belief was that human life was capable of unlimited improvement and it was through education that this improvement was to be attained. Faculty psychology, which suggested human behavior could be improved, was especially appealing to those who held the above belief of the role of education. The school would guide the child in exercising desirable mental faculties, while forcibly suppressing the use of undesirable ones. In this way a free adult, capable of the intelligent and moral judgement necessary to the perfection of a democratic society, could be produced. Since reason was recognized as one of the thirty-seven faculties, arithmetic became an important part of the school curriculum.

Methodology of the day was discipline-based, as orderly behavior was part of the central educational goal. Mastery of fundamentals became the most valued intellectual goal of the elementary schools and resulted in rote instruction and drill in the fundamental operations of arithmetic. The main aims of such instruction were speed and accuracy of computation. During this period arithmetic came to dominate the school day as teachers attempted to achieve these aims.

The period between 1890 and 1930 was marked by a significant reaction to the treatment of education as a formal discipline, as had been done in the past. Curriculum, under the influence of Dewey, became more pragmatic and moved away from mind-exercising formalism. He saw the school as an important tool for social reconstruction. Problems and exercises were based on practical situations.

During this period the stimulus-response psychology, as advocated by Thorndike, replaced the faculty psychology as the basis for research and textbook writing.

During this transition period the teaching of arithmetic was dominated by misconceptions about the new ideas. For example, Thorndike's Law of Exercise simply perpetuated the familiar rote and drill style of instruction. Teachers were hampered by the fact that teacher competency began to be measured by the extent to which pupils secured mastery in speed and accuracy of computation. These results were usually measured by speed achievement tests. It can be seen then, that although advances in the theory of psychology of learning and curriculum reform attempted to make learning a more meaningful experience (in this case practical) for children, the effect upon the classroom situation was not immediately apparent.

The period 1930 to 1950 saw great changes in the area of psychology as it related to education. The stimulus-response theory of learning was pushed aside by the rise of two new theories. One was the Gestalt or Field Theory and the other the psychology of Piaget.

During this period also the controversy over rote versus meaningful instruction was settled. Studies were completed which suggested that the child's ability to transfer, generalize and retain arithmetic was greater when children were given meaningful instruction as opposed to instruction by rote processes. DeVault and Kriewall (1969) state:

In 1935, the official endorsement to meaningful instruction given by the National Council of Teachers of Mathematics (NCTM) became the unchallenged guide to arithmetic instruction well into the 1950's. (p. 17)

The second world war prevented an immediate change in teaching practice towards the adoption of the meaning method until well into the 1950's. Subsequently the great change was rapid and unexpected.

The earliest research leading to the development of the

meaning method was largely psychological research. Judd (1908) carried out a learning experiment which involved striking an underwater target with a dart. Judd took two groups. To one group he explained the refraction principle but did not mention it to the other. The group which had the benefit of understanding the refraction principle far excelled the second group in performance, and the conclusion drawn by Judd was that learning (hitting the underwater target) was strongly influenced by the understanding of principles.

Ebbinghaus, in his early studies of the learning process noted that meaningful material is learned or memorized more rapidly than meaningless or nonsense material. This led to the invention of nonsense syllables for use in his study of memory (DeCecco, 1968).

Tyler (1934) conducted a study of how much course content was retained one year after the completion of a course in Zoology. He found that students did far better on questions requiring the use of concepts presented in the course compared to the results on questions dealing with the recall of unrelated facts. The conclusion drawn was that the teaching of concepts that relate facts in a meaningful manner results in greater retention.

A more recent study by Underwood (1964) showed that among all the conditions of verbal learning, meaningfulness has the strongest influence on the rate of learning.

DeCecco (1968) in a summary of the effect of meaningfulness on verbal learning concludes:

Meaningfulness influences both learning and retention.
The higher the meaningfulness the more rapid the learning
and the longer the materials are retained. (p. 339)

The results from psychological research concerning the effectiveness

of a meaningful approach to teaching were reinforced by educational research in the field of mathematics. The first such important study was completed by Thiele (1938). His study, with second grade children showed that those taught under the generalization method far excelled in performance children taught under the drill method. This was the first major study lending support to the meaningful approach to arithmetic instruction.

Brownell and Moser (1949) conducted a study comparing meaningful and drill methods of instruction in the teaching of subtraction to grade three pupils. Their findings suggested that instruction emphasizing mathematical understandings produced results superior to those based on the drill method.

Miller (1957) reported the Los Angeles study in which the "rule" or drill method and the meaning method were compared. Conclusions drawn as a result of this study were that the meaning method was more effective in the area of computation of fractions, decimals and percentages; the meaning method was effective in establishing retention in processes of computation as well as the understanding of the principles of arithmetic; the meaning method was more effective for the comprehension of complex analysis in arithmetic, indicating a potential superiority for difficult concepts.

An experiment reported by Stokes (1958) compared a method of instruction emphasizing meanings with standard textbook procedures. Findings showed that pupils taught by the meaning method were, on the average, one year higher in achievement than those pupils studying the standard text. He concluded that meaningful instruction will improve learning in arithmetic.

Krich (1964) contrasted meaningful teaching of division of a fraction by a fraction, with teaching using a drill method which provided the child with a rule. Though post-tests did not produce significant differences, a retention test indicated that the meaning method was superior for middle and high I.Q. students.

A study by Shuster and Pigge (1965) compared three methods of teaching the addition and subtraction of fractions in the fifth grade. The treatments used different ratios of time spent on mathematically meaningful or socially significant experiences, and on drill activities. Delayed recall tests indicated that the greater part of class time should be spent on mathematically meaningful or socially significant experiences when the addition and subtraction of fractions was being taught.

The meaningful theory of teaching arithmetic has now been generally accepted by mathematics educators for some time. The first published acceptance of this theory appeared in the tenth yearbook of the National Council of Teachers of Mathematics Yearbook in 1935. Brownell, writing in this yearbook, presented a comparison of the three theories of arithmetic instruction in use at that time - the rote theory, the incidental theory and the meaning theory. In proposing an acceptance of the meaning theory, he concluded his article by stating:

The basic tenet in the proposed instructional reorganization is to make arithmetic less a challenge to the pupil's memory and more a challenge to his intelligence. (p. 31)

The acceptance of the meaning method of arithmetic instruction by Weaver (1950), McSwain (1950), Hildreth (1959), Gibb (1959), may be summarized by this statement made by Van Engen (1955):

It is recognized today that understandings, insights and meanings help children to think about quantity and

to produce rapid and accurate computers. The days of /
 "this is the way you get the answer" teaching are passe.
 Today the teacher encourages the child to structure
 ideas and symbolize them so as to produce the arithmetic
 power that can only develop through meanings and insight.
 (p. 133)

This acceptance has carried on through the 1960's as indicated by Biggs (1963) who stated that the trend at that time was towards the discovery method of teaching as opposed to mechanical procedures with no understanding.

Shulman (1970) stated that the disagreement in mathematics education today centers around discovery versus expository teaching. There is a general acceptance of the meaning approach. The debate now concerns which method presents material in a more meaningful manner.

Summarizing, it is evident that there is general acceptance of the meaning method of arithmetic instruction which is a consequence of research results from the field of psychology, reinforced by studies in mathematics education.

THE IMPORTANCE OF MEASURING MEANINGS

Evaluation of mathematics programs is considerably more important today than was the case a decade ago. Radical changes in the curriculum at all grade levels, coupled with a variety of new pedagogical techniques, have resulted in a much greater variety of textbooks and textbook sequences than we have ever had before in this country. School officials responsible for decisions on the choice of textbooks and educational procedures are asking for more and better information and more powerful ways of evaluating alternatives among which they must choose. (Begle and Wilson 1970, P. 367)

One of the fundamental principles of such an evaluation program is to base all evaluative procedure on the objectives of instruction. This principle has been recognized by educators for some time - for example by Spitzer (1948), Merwin (1961), Sobel and Johnson (1961), Epstein (1968) and Weaver (1970). Since the teaching of meanings and under-

standings in arithmetic has been recognized as an essential part of a modern elementary mathematics program, it would seem imperative that a good evaluation or testing program would include a measure of children's understandings of meanings in arithmetic.

Douglas and Spitzer (1946) suggested several reasons for measuring children's understandings of arithmetical meanings. The first is the effect such measures have on instructional procedures. If an overall evaluation of a mathematics program tests only factual knowledge (computational skills) then teachers, in the belief that they are also being rated by these test results, will teach only for competency in computational skill and the instructional program will suffer.

The effect of evaluating arithmetical meanings upon learning procedures is given as a second reason. They claim that children will learn only that on which they will be tested. It is fruitless to try and encourage children to understand what they are asked to learn, unless their understanding is measured.

The third reason stated for measuring understandings is the effect such measures have on results in research. The worth of experimental programs is almost invariably appraised in terms of test data. Unless all aspects of the experimental program are assessed it is impossible to make valid judgements about such programs.

The authors concluded this section of their article by suggesting that the development of means to judge understandings more adequately will be an important aspect of educational research for some time.

Eads (1959) suggested again that what pupils learn in school

is largely determined by what teachers hold to be important. Children assume that that which is tested is important. The implication is obvious. If speed and accuracy in computations are all that is held to be important in an arithmetic program, then they are all that should be tested. However, if children's understandings are held to be important, then they also should be tested.

Madaus (1961) stated that the mathematics curriculum has been changing gradually, moving away from the position where great importance was attached to only computational skill. This change in curriculum implies that changes are needed in evaluation. We can no longer merely evaluate computational skill. He further suggests that this change in evaluation procedure will improve instruction, improve progress of individual students and aid administrative decisions about curriculum.

The feelings of mathematics educators on the need for assessing children's understandings in arithmetic are contained in the following statement by Koenker (1960):

If we are to teach arithmetic by the meaningful method then it is also necessary to test for meanings and understandings in arithmetic rather than computational ability alone. Testing for meanings and understandings will help evaluate our teaching procedures in a more valid manner, since a valid test is designed to measure the intended outcome of instruction; and who would deny the acquisition of meanings and understandings as a cardinal objective or outcome of a good arithmetic program. (p. 93)

It is apparent that, in view of the general acceptance of the meaning method of arithmetic instruction, evaluating children's understandings of arithmetic would now be an essential part of any good mathematics program. This implies that there are available instruments

suitable for such a testing purpose. The literature would indicate that an abundance of such instruments is not readily available.

Rappaport (1959) stated that the great progress made over the past twenty years in teaching meaning has unfortunately not been accompanied by an adequate evaluative instrument. He further stated that attempts to construct such tests have been left to doctoral candidates and that these need to be improved.

The situation does not seem to have improved greatly. Gray (1966) claimed that the recent curriculum developments in arithmetic have refocussed attention on problems of evaluation. There just did not seem to be adequate instruments for measuring understanding of mathematics.

Ashlock (1968) stated that although mathematics educators have, for thirty-five years, been placing increasing emphasis on the teaching of mathematical understandings, this has not resulted in changes in mathematics achievement tests which remain computationally orientated.

Rea and Reys (1970) state that teachers, attempting to apply the instructional schemes of modern mathematics, are handicapped by a lack of adequate instruments for measuring students' abilities to grasp the instructional objectives of these programs.

In summary, the literature indicates that the assessment of children's understandings of arithmetic is essential, but that there are very few instruments available that give an accurate measure of these understandings.

ATTEMPTS TO EVALUATE MEANING

The first major attempt to construct a test of mathematical understandings was the result of a pilot study by Glennon (1949). He analyzed all arithmetic tests published since 1915 and found relatively few items designed to measure understandings. His test consisted of eighty multiple choice items directed towards the understanding of the decimal system of notation, integers and processes, fractions and processes, and decimals and processes. He intended the test for use with students at or above the seventh grade, since he felt this was the lowest grade level at which all these principles could reasonably have been taught.

Johnson and Trimble (1954) have given a few examples of exercises that test meanings and understandings as guides to teachers who would like to construct such tests.

The Board of Education of the City of New York, realizing the need of teachers to determine how adequately pupils were gaining mathematical understandings, had a test constructed under the direction of its Bureau of Educational Research. Level I of the test, primarily for use in grade 1, included items designed to measure understanding of one-to-one correspondence, the cardinal number concept, and the ordinal number concept. Level 2, primarily for use in grade 2, also included items to measure an understanding of the ordinal number concept. Wrightstone (1956) and his associates in the construction of this test did not include items which would measure an understanding of place value or the properties which hold for operations on whole numbers.

A seventy-two item test, which tested the understanding of sixteen properties of the number system was designed by Rappaport (1958).

This was a three part test aimed towards students in grades 7 and 8.

Dutton (1964) prepared multiple choice tests for each of grades 3 through 6. An attempt was made to include the major mathematical understandings presented at each grade level, therefore these tests involved a greater number of different understandings than do many of the tests which have been developed. As the author expected the tests to be used diagnostically no norms were established.

Edwards (1965) developed a multiple choice test of understandings for use in grades 3 through 6. The test was designed to measure children's understandings of place value, the ordinal concept, reading numerals and properties of the basic operations for whole numbers. This was apparently the first test for use in the intermediate grades which was designed to measure understandings of these properties of the basic operations.

A three-part test was constructed by Ashlock (1966). The test consisted of fifty items measuring such basic principles as one-to-one correspondence, cardinal number concept, ordinal number concept, place value and properties of the basic operations with whole numbers. The test was intended for use in the primary grades.

Flournoy (1968) reports on the construction of two tests aimed at measuring children's understandings of arithmetic principles. Ninety-four principles were determined and multiple choice items were designed so written computation would not be needed and responses could be selected by using principles. The primary test consists of one hundred fourteen items and the intermediate test one hundred thirty-two items.

A test constructed at the University of Wisconsin, The Wisconsin Contemporary Test of Elementary Mathematics (DeVault et al,

1967), was prepared as a measure of the mathematics taught in contemporary mathematics programs. The test has two forms at each of two levels - grade 3 - 4 and grade 5 - 6. The grade 3 - 4 test includes fifty-two items and the grade 5 - 6 test, sixty items. There are three basic subtests included in each test, fundamentals, operations, measurement and geometry. Each of these subtests is divided into two parts. Items in the first part test facts; those in the second part concepts.

STANDARDIZED TESTS USES AND SOME CRITERIA FOR SELECTION

Previous sections in this chapter have attempted to show, firstly, that the meaning method of arithmetic instruction had a sound basis in psychology and was generally accepted by mathematics educators; secondly, to illustrate that in view of the general acceptance of the meaning method, children's understandings of these meanings should be evaluated; thirdly, present attempts that have been made to evaluate these understandings. In this section the role of standardized tests in a complete evaluation program will be presented as well as criteria on which to base the selection of a standardized test.

Brueckner (1959) stated that the chief contributions of testing and evaluation to arithmetic instruction are (1) the selection and clarification of objectives which serve as guides for testing and instruction, (2) the determination of the rate of growth and the progress made by each learner in achieving accepted objectives, (3) the provision of a basis on which teachers can set up educational experiences adapted to the needs, interests and ability of the learner, (4) motivation and guidance of learning, especially by helping children to evaluate their own responses and behavior, (5) the location, diagnosis

and treatment of learning difficulties, (6) the basis for coordinating improvement programs in related fields such as arithmetic, reading, science and social studies. Furthermore, he advocates the use of both standardized and non-standardized instruments.

Several opinions on the use of standardized tests in the overall evaluation of the mathematics program will be presented. DeVault and Kriewall (1969) claim that standardized tests can be used to (1) assess the progress of each child and evaluate this progress in terms of the expectations for the individual, (2) to recognize weaknesses in three different areas. Teachers can recognize weaknesses in instruction by comparing norms to class means. Principals can recognize weak teachers by comparing class results to the expectations for that class. Administrators can be helped in the recognition of weak schools. (3) System-wide evaluation makes it possible to examine the mathematics program in the total system as compared to national norms.

Ahmann, Glock and Wardberg (1960) state that standardized tests can be used to (1) provide teachers with a criteria for checking their own emphasis in teaching, (2) determine pupil progress, (3) compare achievement between various subject matter areas and specific phases of a particular area, (4) diagnose difficulties in achievements, (5) to help group children.

In the opinion of Greene, Jorgensen and Gerberich (1953), standardized tests provide a class analysis and diagnosis which help the teacher to plan instruction, enable schools to identify the special abilities of their pupils and to challenge these children to greater efforts and aid the school administrators in determining pupil gradation and placement.

Since this research is concerned with standardized achievement tests it seems appropriate to present views on the use of this particular type of test. Mehrens and Lehman (1969) state that standardized achievement tests assess pupil's knowledge and skills at a particular time. They further state that they should be used in conjunction with other evaluative devices for making decisions regarding (1) the assignment of course grades, (2) vocational and educational guidance, (3) promotion, (4) placement, (5) teacher evaluation, (6) instruction, (7) research.

Garrett (1959) feels that achievement tests are useful for survey purposes, i.e. to determine a class standing in relation to some norm and for guidance and evaluation; i.e. to provide a clearer understanding of what individual pupils have learned or have failed to learn and as a result remedial work may be prepared and instruction improved.

Standardized tests are in wide use, particularly in the area of curriculum evaluation. This statement is supported by the findings of Welch (1969) who found that standardized tests were the second most popular instrument used for this purpose. They are second only to instruments designed expressly for the purpose of evaluating a particular curriculum.

Reasons for this wide use are presented by DeVault and Kriewall (1969) who state (1) they (standardized tests) have been carefully prepared, (2) their validity and reliability have been carefully determined, (3) alternate forms of the test are available and (4) norms have been made available.

Garrett (1959) states that standardized tests are superior to routine exams as they are (1) better planned in that they present a

consensus of opinion of what should be tested in a particular area, (2) more objective, (3) constructed to more exact specifications.

Lindvall (1967) identifies six criteria to be used in the selection of a standardized test. These criteria are validity, reliability, adequacy of the norms, ease of administration, ease of scoring, economy. The validity of a test refers to the extent to which a test measures the qualities, abilities and skills that it is supposed to measure. In achievement tests content validity is of prime importance (Gronlund 1968, p. 106; Ahman, Glock and Wardeberg 1960). Reliability refers to the consistency of results obtained from the test. A highly reliable test, if administered to the same group of subjects on different occasions, would give results that varied only to a small degree. Achievement tests should have reliabilities in the range .80 - .95 (Dutton 1964). Adequacy of the norms refers to the representativeness of the sample used in establishing the norms. Norms established by using only a small sample or a sample from one local area would not be as useful for comparison purposes as norms established by using a large sample from a wide geographic area. Ease of administration refers to the amount of special training needed, both by the person administering the test and those writing, in order to administer the test. A test which can be administered only by a highly trained person is not as useful as a test that can be administered successfully by a person with very little training as these specially trained people are not readily available to all schools. Ease of scoring refers explicitly to the ease of the marking of the test, and implicitly to the objectivity of the test items. Tests with well designed multiple-choice items are both easy to score, either by hand-scoring or by machine-scoring, and

highly objective as no marker judgement is required in the scoring. Economy refers to the cost of the test. The selected test should have the lowest per pupil cost in terms of administration and scoring, all other things being equal. However, economy is certainly not one of the major criteria to consider in selecting a test, for a cheap test conceivably may be quite useless.

These criteria are the generally accepted criteria to be used in the selection of standardized test. Cronbach (1960) lists validity, reliability, ease of application, cost, equivalent or comparable forms, scoring method and time required for testing as a basis for test selection.

The Otis Score Card for Rating Standardized Tests (Greene and Jorgenson 1940, p. 121) is divided into eight categories by which a standardized test may be evaluated. They are the manual, validity, reliability, reputation, ease of administration, ease of scoring, ease of interpretation, typography and make-up.

A more recent listing of criteria for test selection are those put out by the Center for the Study of Evaluation, Los Angeles, California (Horpfner, 1970). This listing is published under the title "The Mean Evaluation Criteria". The word "mean" is derived from the four major categories used in this test evaluation procedure, measurement validity, examinee appropriateness, administrative usability and normed technical excellence. Each of these main categories is broken down into from between two and ten sub-categories. Measurement validity has two sub-categories; examinee appropriateness, six; administrative usability has ten and normed technical excellence, six.

In evaluating a test each sub-category is assigned a score.

The maximum score for each sub-category varies from one to ten. The total for each major category is fifteen. Each major category is then given one of three ratings, good, fair or poor. A rating of "good" (12 - 15 pts.) indicates that the test meets the criterion very well. A "fair" (8 - 11 pts.) indicates that the instrument is probably among the better tests available, but it does not meet well the criteria desired. A test that is unsatisfactory in that specific criterion receives a "poor" rating (0 - 7 pts.).

SUMMARY

In summary, the literature appears to give support to the contention that the meaningful teaching of arithmetic has long been an acceptable educational objective. This acceptance has the backing of research in both the field of psychology and educational research. As such, its evaluation is an essential part of any arithmetic program in order to stress its importance to both teachers and students. Standardized tests can play an important role in such evaluation if used properly in conjunction with other evaluative techniques.

The last section of this chapter details criteria to be used in the selection of a standardized test. Those generally accepted by experts in the field of evaluation are validity, reliability, ease of administration, ease of scoring and economy. A detailed examination of any other list of criteria would reveal that these five are always included.

Chapter 3

METHODS AND PROCEDURE

THE SAMPLE

The population for this study consisted of all pupils in grades one and two in six schools of the Edmonton Separate School Board. Three of the schools (St. Peters, St. Martin, St. Agnes) were following the mathematics programs set down in the STA1 and STA2 textbooks. The other three schools (St. Jerome, St. Stanislaus, St. Boniface) were following the mathematics programs set down in the AW1 and AW2 textbooks. The schools used in this study were selected by the Board.

The sample was obtained by having four schools (St. Martin, St. Agnes, St. Boniface, St. Stanislaus) select twelve children from each of grades one and two and two schools (St. Peter, St. Jerome) select six children from each grade. In selecting the children the schools were asked to select equal numbers in each grade in each of the three achievement groups, high, average and low. The sample, thus, consisted of one hundred and twenty children, thirty in each of the four mathematics programs, AW1, AW2, STA1 and STA2. Each group of thirty contains ten pupils in the high achievement group, ten in the average achievement group and ten in the low achievement group.

During the statistical analysis the two grade one groups and the two grade two groups were compared using I.Q. as a covariate. These results are summarized in Table 1. This analysis showed that the AW1 group and the STA1 group could be considered to be equal in terms of I.Q.

The AW2 group and the STA2 group were found to be similarly equal.

The ability groups of the AW1 and AW2 program were compared to those of the STA1 and STA2 programs respectively on the basis of I.Q. Results showed that the high, average and low ability groups of the two AW grades were not significantly different from the high, average and low ability groups of the corresponding STA grade. These results are presented in Table 2.

THE INSTRUMENTS

The Cooperative Primary Test, Mathematics, Form 12A

This is a fifty-five item test designed to be administered to grade one children in the spring and grade two children in the fall. The test handbook identified eight categories or sub-tests of the whole test. These are, number, symbolism, operation, function and relation, approximation, measurement, estimation and geometry. The number of items in each category ranges from a high of eighteen in the number category to a low of one in both the approximation category and the estimation category. This researcher identified seven sub-tests of the Cooperative Test in order to have the sub-tests more closely coincide with the categories of objectives identified for each program (see Appendix B). The sub-tests identified by the researcher are number-numeration, basic facts, time, money, measurement, fractions and geometry. The number of items in each of these sub-tests range from a high of nineteen on the number-numeration sub-test to a low of two on the time sub-test (see Table 3).

The test handbook is divided into two main sections and two appendices. The first section entitled "Manual" contains information regarding the staff responsible for the test, test philosophy and

Table 1

ANALYSIS OF VARIANCE COMPARING BY GRADE LEVEL MEAN I.Q. SCORES
OF THE AW AND STA GROUPS

SOURCE OF COMPARISON	MEAN SCORE AW	STA	SS	DF	MS	F	P
AW1 - STA1	113.3	113.4	0.0	1	0.0	0.0	1.00
AW2 - STA2	110.1	110.5	21.25	1	21.25	0.03	0.873

Table 2

ANALYSIS OF VARIANCE COMPARING BY GRADE LEVEL MEAN I.Q. SCORES
THE THREE ABILITY GROUPS OF THE AW PROGRAM TO THE THREE ABILITY GROUPS OF THE STA PROGRAM

SOURCE OF COMPARISON	MEAN SCORE AW	STA	SS	DF	MS	F	P
High AW1 - High STA1	119.2	115.5	68.44	1	68.44	0.845	0.370
Average AW1 - Average STA1	113.0	116.8	45.00	1	45.00	2.92	0.105
Low AW1 - Low STA1	107.6	107.9	0.375	1	0.375	0.006	0.940
High AW2 - High STA2	116.2	119.9	68.44	1	68.44	1.14	0.301
Average AW2 - Average STA2	106.7	112.3	156.75	1	156.75	1.66	0.213
Low AW2 - Low STA2	107.4	99.3	328.00	1	328.00	3.59	0.074

development, a statement of what the instrument purports to test, and a section on interpreting test results. The purpose of the test as stated by the handbook is to test mathematics, not just number work or straight computation.

The section on interpreting test results explains in detail how to interpret the results of the test and then how to use them. This information is provided specifically for two groups, firstly program evaluators and curriculum planners, and secondly, teachers.

The second section of the handbook, entitled "Technical Report" contains information regarding pretesting administrations, norming and equating, the norms and statistical characteristics of the final forms of the test including information regarding the tests validity, reliability, difficulty.

Appendix A contains directions for administering the test and Appendix B contains directions for scoring the test.

The test will now be discussed in terms of the criteria to be used in the selection of a standardized test stated in Chapter 2.

Validity

The handbook for the Cooperative Test claims that the test will focus on skills and concepts basic to future development in mathematics and that the tests are designed to measure attainment of major educational objectives, regardless of particular curriculum program and methods. It is further suggested in the handbook that each test user make an individual judgment of content validity with respect to his own instructional practices and educational aims.

The content validity of the Cooperative Test was assessed with reference to the objectives of each of the four mathematics programs

in question. The procedure used in this assessment will be described later in this chapter.

It was found that the Cooperative Test had a content validity coefficient of .53 for the AW1 program - i.e. 53% of the objectives identified for that program were tested, .57 for the STA1 program, .56 for the AW2 program and .42 for the STA2 program.

In conjunction to the assessment of the content validity of the Cooperative Test for each of the four mathematics programs used in this study, the number of valid items on the test for each of the four programs was identified. Of the fifty-five items contained on the Cooperative Test, thirty-six were judged to be valid items for the AW1 program, twenty-nine for the STA1 program, forty-five for the AW2 program and thirty-five for the STA2 program.

The number of unsuitable items on the test for each program was also determined. An unsuitable item is one which has an Item Reliability Index of less than 0.1. With reference to the AW1 program the Cooperative Test had fourteen unsuitable items, for the STA1 program sixteen unsuitable items, for the AW2 program seventeen unsuitable items, and for the STA2 program nine unsuitable items.

Reliability

The internal consistency coefficient presented in the handbook for the Cooperative Test were calculated using the Kuder-Richardson Formula 20. The reliability coefficient for this instrument associated with the grade one sample is .86 and for grade two is .84.

The reliability coefficients for this test associated with the samples used for this study were also calculated using the Kuder-Richardson 20 Formula. For the AW1 group the reliability coefficient

is .85, for the STA1 group .76, for the AW2 group .90, and for the STA2 group .93. The mean reliability coefficient for the grade one sample used in this study is .81. For the grade two sample the mean reliability coefficient is .92. All reliability coefficients, except the one for the STA1 group, fall within the acceptable range identified in Chapter 2, (.80 - .95).

Ease of Administration

The administration of this test requires no special training. The classroom teacher would need only to read through the directions for administering the test in a manner that ensures accurate results. Children indicate their responses by drawing a large X through the answer they think is correct.

One difficulty encountered by the researcher in administering this test was that in some places the items are not spaced as well as they should be, and the children either had difficulty locating the question they were to answer or inadvertently answered two questions at once. This difficulty can probably be explained by both improper spacing of items and the fact that the answer boxes for different items are not of uniform size.

The time required to administer this test is fifty minutes. The handbook recommends that the test be administered in two separate sessions - a suggestion readily endorsed by the researcher.

Ease of Scoring

The Cooperative Test can be easily handscored using the key supplied with the test. Directions are given in the handbook for interpreting the scores. Norms are supplied in the handbook for comparison purposes. The difficulty index of this test was calculated for each of

the four programs used in this study by averaging the difficulty index for each item. For the AW1 group the difficulty index is .67, for the STA1 group .64, for the AW2 group .86 and for the STA2 group .78. The mean difficulty index for the grade one sample used in the study is .66, and for the grade two sample .82. The index for the grade one sample would indicate that the test would discriminate grade one pupils to a reasonable degree. However, the grade two index indicates a general lack of discrimination which indicates the test should be used only in the fall with grade two students as was the intention of the test designers.

Economy

The price of this test is \$3.25 per 20 tests. When assessing the economy of this test one must consider the price in conjunction with the qualities discussed above in order to determine the value of the test for the required purpose per dollar spent.

An evaluation of the Cooperative Test by the Center for the Study of Evaluation at U.C.L.A. rated the instrument as follows. Under the heading "Measurement Validity" a score of eight out of a possible score of fifteen resulted in a "fair" rating. A score of seven out of fifteen on "Examinee Appropriateness" resulted in a "poor" rating. A rating of "good" on "Administrative Usability" resulted from a score of thirteen out of fifteen. The category "Normed Technical Excellence" received a "fair" rating from a score of nine out of fifteen. In averaging these four ratings, one arrives at a general rating of "fair" for the Cooperative Test. A rating of "fair" indicates the instrument is among the better tests available but that results from a test with such a rating should be interpreted more cautiously than a test with a "good" rating.

The Modern Mathematics Understanding Test, Primary Form C (The SRA Test)

This is a forty-two item test designed to be administered to grades one and two. Seven sub-tests were identified by the researcher. They are number-numeration, basic facts, fractions, time, money, measurement and geometry. As shown in Table 3, the number of items in each sub-test range from a high of twenty-two in the basic facts sub-test to a low of one in the fractions sub-test.

There is no test handbook. The test results returned as a result of the machine marking plan (Plan 17) used in this study lists a stanine score for each pupil which indicates the position of the pupils score in comparison to the entire grade level population to which the pupil belongs.

In discussing the instrument in terms of the criteria identified in Chapter 2, only the results obtained by the researcher in using the tests will be mentioned as no other statistical information is available on the SRA Test.

Validity

The assessment of the content validity of the SRA Test with reference to the four mathematics programs used in this study gave rise to the following results. The content validity coefficient of the SRA Test with reference to the objectives of the AW1 program was found to be .50; for the STA1 .20; for the AW2 program .44, and for the STA 2 program .40.

The number of items on the SRA Test judged to be valid for the AW1 program was twenty-four of forty-two; for the STA1 program seven of forty-two; for the AW2 program thirty-three of forty-two, and for the STA program twenty-three of forty two.

The number items on the SRA Test judged to be not suitable for the AW1 program was eleven of forty-two; for the STA1 program nineteen of forty-two; for the AW2 program four of forty-two, and for the STA2 program three of forty-two.

Reliability

The reliability coefficients for this test associated with the samples used in this study were calculated using the Kuder-Richardson 20 Formula. They are, for the AW1 program .82; for the STA1 program .61; for the AW2 program .91, and for the STA2 program .93. The mean reliability coefficient for the grade one sample is .72 and for the grade two sample .92. The coefficient for the grade one sample falls outside the acceptable range due to the very low reliability coefficient associated with the STA1 program. The coefficient associated with the grade two sample is well within the acceptable range.

Ease of Administration

The administration of this test also requires no special preparation other than carefully reading the instructions to be given the children prior to administering the test.

The items are well spaced and children seemed to have little difficulty locating the correct question and indicating their response. Responses were indicated by filling in a small circle beneath the answer thought to be correct. This researcher was a little apprehensive of grade one children's ability to fill in these circles properly, but they seemed to have little trouble in accomplishing this task.

The time required for writing this test is sixty minutes.

Ease of Scoring

This test could be easily hand-scored. However, it can be machine-scored. The machine-scoring must be done by SRA, which requires

the tests be sent to Chicago. The Modern Mathematics Understanding Test (the SRA Test) may not be ordered without also ordering the marking service.

The difficulty index for this instrument associated with pupils in the AW1 program was .414; with the STA1 program .37; with the AW2 program .67, and with the STA2 program .58. The mean difficulty index for the grade one sample is .39, and for the grade two sample .62. The desired difficulty index for a test is .50. This gives a test with maximum discrimination (Granlund 1968, Adkins Wood 1960).

Economy

The SRA Test may be purchased for various rates depending upon the type of marking service required. As mentioned before, the tests may not be purchased without purchasing the marking service. The rates vary from \$1.07 per test to \$.65 per test.

PROCEDURE

Determining Content Validity

Ebel (1956) states that the simplest and most direct method of obtaining evidence of content validity is to have the test examined by a competent judge. He goes on to say that the content validity of a test is determined by its relevance to the objectives of instruction rather than by its coverage of the materials of instruction.

The objectives of each of the four mathematics programs used in this study were determined by the researcher. The Teacher's Guides for each program lists instructional objectives for each division of the textbook. These objectives were combined to form the objectives that

Table 3

THE NUMBER OF ITEMS ON EACH OF THE VARIOUS SUB-TESTS
IDENTIFIED BY THE RESEARCHER OF
(a) THE COOPERATIVE TEST
(b) THE SRA TEST

SUB-TEST	COOPERATIVE TEST	SRA TEST
Basic Facts	17	22
Number-Numeration	19	8
Time	2	2
Money	3	2
Measurement	5	2
Fractions	3	1
Geometry	6	5

Table 4

NUMERICAL DESCRIPTION OF THE INSTRUMENTS USED IN THIS STUDY:
 THE COOPERATIVE PRIMARY TEST, MATHEMATICS, FORM 12A
 THE SRA MODERN MATHEMATICS UNDERSTANDING TEST, PRIMARY FORM C

	Cooperative	SRA
No. of Items	55	42
Time Required	50 mins.	60 mins.
Content Validity		
AW1	.53	.50
STA1	.57	.20
AW2	.56	.44
STA2	.42	.40
Valid Items		
AW1	36	24
STA1	29	7
AW2	45	33
STA2	35	23
Unsuitable Items		
AW1	14	11
STA1	16	19
AW2	17	4
STA2	9	3
Reliability		
(a) Handbook		
Grade 1 spring	.86	
Grade 2 fall	.84	
(b) Tested		
AW1	.84	.82
STA1	.76	.61
AW2	.90	.91
STA2	.93	.93
Difficulty		
AW1	.67	.41
STA1	.64	.37
AW2	.86	.67
STA2	.78	.58
Price	\$3.25 /20	\$1.07 - \$.65 /1 (includes scoring)

this researcher identified for the four programs. A complete listing of the objectives appears in Appendix B.

The examination of the tests was performed by a committee of four judges including the researcher (see Appendix A). The examination was a matching process whereby a test item was matched to the particular objective it was judged to test. An objective was defined to be tested if two judges assessed it to be tested by at least one test item. It was not necessary for the judges to agree as to the item which tested a particular objective. A summary of the matching process appears in Appendix C.

Judge reliability was checked by means of the Arrington Formula. (Feifel and Lorge, 1950). This formula measures the consistency of the judges' observations (in this case matching item to objective) and gives a reliability coefficient based on the formula $r = 2a/2a+d$ where "a" is the total number of agreements between the judges and "d" the total number of disagreements.

Four content validity coefficients were calculated for each test; one with reference to the objectives of each of the four mathematics programs used in this study. The coefficient is the ratio of objectives tested to the total number of objectives.

As a result of this matching process the number of valid items on each test for each of the four programs were identified. An item was called valid if two judges assessed it to test some objective of the program in question. It was not necessary that the judges be in agreement on the objective tested by the item.

The number of unsuitable items on each test for each mathematics program was ascertained by counting the number of items

that had an Item Reliability Index of less than 0.1. This index is part of the item analysis printed out by computer program Test 04.* This meant that eight separate counts of unsuitable items were made, four for each test.

Administering the Tests

The tests were administered by the researcher between April 26, 1971 and May 5, 1971. The schools were matched in pairs, that is, a school on the AW program was paired with a school on the STA program. Two pairs were given the Cooperative Test first and one pair of schools the SRA Test first. The two tests were administered on separate days. At four schools the tests were administered on consecutive days and at two schools on alternate days at the request of the school. The tests were administered to both grades at the same sitting.

Preliminary to administering the actual test the children were taught the method of responding to the test items by means of a ten item pilot test. The tests were then administered in two parts with a five minute break between parts.

The children's responses were transferred to data cards and marked using computer program Test 01. An item analysis of each test was completed. The individual sub-tests of each test were marked by the researcher.

STATISTICAL PROCEDURES

For all hypotheses except hypotheses one and twelve, the

*The computer programs referred to in this study (e.g. Test 04) are those maintained by the Division of Educational Research, University of Alberta, Edmonton.

analysis of variance technique was chosen as the statistical means of analyzing the data in this research.

Hypothesis one, which was looking for significant differences in the content validity of the two tests used in this research with reference to a particular mathematics program was analyzed by using a test to compare the significance of the difference between two independent proportions (Ferguson, p.176).

Hypothesis twelve, which was looking for significant differences between the number of items judged not suitable on each of the two tests with reference to one of the four mathematics programs, was analyzed using the same technique as for hypothesis one.

Hypotheses two and three, which were looking for significant differences between achievement scores of pupils in different mathematics programs on a particular test were analyzed by means of a two-way analysis of covariance (ANCOV20) using I.Q. as a covariate.

Hypothesis four compared the achievement scores of each of the four mathematics on each of the two tests used in this study. The hypothesis was analyzed using a one-way analysis of variance (ANOVA10).

Hypotheses five and six, which looked at the differences in the achievement scores of the three ability groups in the AW program at each grade level as compared to the ability groups in the STA program, were analyzed by means of one-way analysis of covariance (ANCOVA10).

One-way analysis of variance (ANOVA10) was used to analyze hypotheses seven, eight, nine and ten. These hypotheses looked at differences in the achievement scores of pupils in the AW program and pupils in the STA program on the two major sub-tests of the instruments

used in this study.

Hypothesis eleven which compared the achievement scores of pupils using the machine markable format of the SRA Test with those using a hand scoring format was analysed by means of a one-way analysis of covariance (ANCOVA).

INTELLIGENCE AS A COVARIATE

The pupils used in this study were selected by the schools. Equal numbers of pupils were chosen by each school of the three ability groups defined in Chapter 1. The basis for group inclusion was mathematics achievement as determined by the schools.

In order to have a basis for comparing the groups of children, I.Q. was chosen as a covariate as it was felt to be impossible to compare individual achievement scores. To partially justify the selection of I.Q. as a covariate, Pearson product-moment correlations were computed between pupil I.Q. and pupil achievement scores on the two tests. These correlations were computed for the whole grade one sample; the whole grade two sample and each of the samples from the four mathematics programs.

The results as shown in Table 4 indicate a significant correlation between pupil I.Q. and achievement score in nine of the twelve cases. The correlation between I.Q. and achievement score was not significant for pupils in the AW2 program and the Cooperative Test and the AW1 pupils on both tests. All other correlations were significant at or beyond the .01 level except that associated with the STAl pupils on the SRA Test which was significant at the .05 level.

These results seem to justify the fact that comparing the

groups on the basis of I.Q. may be considered equivalent to comparing them on the basis of mathematics achievement.

Table 5

CORRELATION BETWEEN I.Q. AND ACHIEVEMENT SCORE
FOR THE PUPILS USED IN THIS STUDY

Sample	<u>Correlation</u>	
	Cooperative Test	SRA Test
Grade 1	.395 *	.367 *
Grade 2	.500 *	.524 *
AW1	.313	.325
STA1	.487 *	.445 **
AW2	.248	.472 *
STA2	.705 *	.603 *

* Significant at the .01 level

** Significant at the .05 level

Chapter 4

THE FINDINGS OF THE STUDY

The findings of the study are presented in three sections. The first section deals with the content validity of the two tests used in this study with reference to the objectives of the four mathematics programs against which the tests were matched. The second section contains the analysis of the data with respect to the hypothesis stated in Chapter 1 and further described in Chapter 3. The third section contains an analysis of data not directly related to the stated hypotheses.

THE FINDINGS WITH RESEPECT TO CONTENT VALIDITY

The Cooperative Test and the AW1 Program

As reported in Chapter 3, the content validity coefficient of the Cooperative Test with respect to the objectives of the AW1 program is .53, (i.e. 53% of the objectives of the AW1 program are tested by the Cooperative Test). Thirty objectives were recognized for the AW1 program (see Appendix B) and of these sixteen were judged to be tested. The judge reliability coefficient associated with the determination of this content validity coefficient is .77.

A breakdown of the objectives tested by this instrument revealed that six of ten objectives listed as number-numeration objectives were judged to be tested and four of nine of those listed as addition and subtraction objectives. Table 6A gives a complete breakdown of the results of the assessment of the content validity.

The Cooperative Test and the STA1 Program

The assessment of the content validity of the Cooperative Test with respect to the STA1 program revealed that seventeen of the thirty objectives identified for this program were judged to be tested. This results in a content validity coefficient of .57. The judge reliability coefficient was .79.

As shown in Table 6B, eight of the fourteen number-numeration objectives of this program were judged to be tested and four of seven of the addition and subtraction objectives. None of the geometry objectives of the STA1 program were tested by this instrument.

The Cooperative Test and the AW2 Program

Associated with the AW2 program the Cooperative Test was found to have a content validity coefficient of .56, i.e. twenty-three of the forty-one objectives identified for this program were judged to be tested. In assessing this content validity coefficient the judge reliability was .74.

Of the thirteen number-numeration objectives identified for this program seven were judged to be tested as were six of the eight addition and subtraction objectives. Of the six geometry objectives identified for this program only one was judged to be tested. Table 6C gives a complete breakdown of the assessment of the content validity of the Cooperative Test for the AW2 program.

The Cooperative Test and the STA2 Program

The content validity coefficient of the Cooperative Test with respect to the STA2 program is .45. This resulted from eighteen of the forty-three objectives identified for this program being judged

tested by the committee of judges. The judge reliability coefficient associated with this assessment is .77.

As shown in Table 6D, six of the ten number-numeration objectives were judged to be tested. Six of the ten addition and subtraction objectives were similarly judged to be tested. No geometry objectives of the eight identified for the STA2 program were judged to be tested. Two of the seven multiplication and division objectives were judged to be tested.

The SRA Test and the AW1 Program

Fifteen of the thirty objectives identified for the AW1 program were judged to be tested by the SRA Test. Thus the content validity coefficient associated with this test for the AW1 program is .50. The judge reliability coefficient associated with this assessment is .81.

A full breakdown of the judge's decisions regarding objectives of the AW1 program tested by this instrument can be found in Table 6E. Six of the ten number-numeration objectives were judged to be tested and six of the nine of the addition and subtraction objectives.

The SRA Test and the STA1 Program

The assessment of the content validity of the SRA Test with respect to the objectives of the STA1 program resulted in six of the thirty objectives being judged tested. The content validity coefficient for the SRA Test in this case then is .20. The judge reliability associated with this assessment is .94.

Four of the fourteen number-numeration objectives were judged to be tested and one of seven addition and subtraction objectives. One of four geometry objectives were judged to be tested but none of the money or measurement objectives were tested. These

Table 6A

RESULTS OF CONTENT VALIDITY ASSESSMENT
COOPERATIVE TEST AND THE AW1 PROGRAM

Objective	1	2	3	4	Tested	Objective	1	2	3	4	Tested
N 1	X	X	X	X	Y	A 6					N
2	X	X	X	X	Y	7		X			N
3		X			N	8		X			N
4		X			N	9			X		N
5	X		X	X	Y	Mo 1	X	X	X	X	Y
6			X	X	Y	2		X	X	X	Y
7	X	X	X	X	Y	3					N
8	X	X	X	X	Y	F 1		X	X		Y
9					N	2			X		N
10					N	T 1	X	X	X	X	Y
A 1		X	X	X	Y	Me 1	X	X	X	X	Y
2	X				N	2					N
3	X	X	X	X	Y	3					N
4	X	X	X	X	Y	G 1		X	X	X	Y
5		X	X	X	Y	2					N

KEY

N - Number Numeration
A - Addition + Subtraction
M - Multiplication + Division
F - Fractions

G - Geometry
Me- Measurement
Mo- Money
T - Time

Content Validity Coefficient .53

Judge Reliability .77

Table 6B

RESULTS OF CONTENT VALIDITY ASSESSMENT
COOPERATIVE TEST AND THE STA1 PROGRAM

Objective	1	2	3	4	Tested	Objective	1	2	3	4	Tested
N 1	X	X		X	Y	A 2			X	X	Y
2		X	X	X	Y	3				X	N
3	X	X	X		Y	4		X	X		Y
4	X	X	X	X	Y	5	X	X		X	Y
5			X		N	6			X		N
6			X		N	7					N
7	X	X	X	X	Y	Mo 1	X	X	X	X	Y
8	X	X	X	X	Y	2	X	X			Y
9					N	3			X	X	Y
10				X	N	4					N
11	X	X	X		Y	G 1			X		N
12	X			X	Y	2					N
13					N	3					N
14				X	N	Me 1			X	X	Y
A 1			X	X	Y	2			X	X	Y

KEY N - Number Numeration
 A - Addition + Subtraction
 M - Multiplication + Division
 F - Fractions

G - Geometry
 Me- Measurement
 Mo- Money
 T - Time

Content Validity Coefficient .57

Judge Reliability .79

Table 6C

RESULTS OF CONTENT VALIDITY ASSESSMENT
COOPERATIVE TEST AND THE AW2 PROGRAM

Objective	1	2	3	4	Tested	Objective	1	2	3	4	Tested
N 1	X	X	X	X	Y	M 1	X	X	X	X	Y
2	X	X	X	X	Y	2		X	X	X	Y
3		X			N	3					N
4	X	X		X	Y	4	X	X		X	Y
5				X	N	5					N
6			X		N	F 1	X	X	X		Y
7	X	X	X	X	Y	2		X	X		Y
8	X	X	X	X	Y	Me 1	X	X	X	X	Y
9	X	X	X	X	Y	2				X	N
10					N	3					N
11				X	N	T 1	X	X	X	X	Y
12	X	X	X	X	Y	Mo 1	X	X	X	X	Y
13				X	N	2			X	X	Y
A 1	X	X	X	X	Y	3					N
2	X	X	X	X	Y	G 1					N
3					N	2					N
4	X	X	X		Y	3					N
5	X	X		X	Y	4		X	X	X	Y
6				X	N	5					N
7	X		X	X	Y	6					N
8		X		X	Y						

KEY N - Number Numeration
 A - Addition + Subtraction
 M - Multiplication + Division
 F - Fractions

 G - Geometry
 Me- Measurement
 Mo- Money
 T - Time

Content Validity Coefficient .56

Judge Reliability .74

Table 6D

RESULTS OF CONTENT VALIDITY ASSESSMENT
COOPERATIVE TEST AND THE STA2 PROGRAM

Objective	1	2	3	4	Tested	Objective	1	2	3	4	Tested
N 1	X	X	X	X	Y	M 3	X		X		Y
2	X	X	X	X	Y	4					N
3					N	5					N
4	X	X	X	X	Y	6	X			X	Y
5					N	7					N
6	X	X	X		Y	Mo 1	X	X	X	X	Y
7	X	X	X	X	Y	2	X	X	X	X	Y
8					N	3					N
9				X	N	4					N
10	X		X	X	Y	G 1					N
A 1	X			X	Y	2					N
2	X	X	X		Y	3					N
3	X	X	X		Y	4					N
4	X			X	Y	5				X	N
5		X		X	Y	6					N
6	X	X		X	Y	7					N
7					N	8					N
8				X	N	Me 1	X	X	X	X	Y
9		X	X		N	2					N
10		X			N	3					N
M 1					N	4					N
2	X		X	X	Y						

KEY N - Number Numeration
 A - Addition + Subtraction
 M - Multiplication + Division
 F - Fractions

G - Geometry
 Me- Measurement
 Mo- Money
 T - Time

Content Validity Coefficient .45

Judge Reliability .77

Table 6E

RESULTS OF CONTENT VALIDITY ASSESSMENT
SRA TEST AND THE AW1 PROGRAM

Objective	1	2	3	4	Tested	Objective	1	2	3	4	Tested
N 1					N	A 6	X	X		X	N
2	X		X		Y	7			X	X	Y
3					N	8	X	X	X	X	Y
4		X	X		Y	9	X	X	X	X	Y
5	X		X	X	Y	Mo 1					N
6	X	X	X	X	Y	2	X	X	X		Y
7	X			X	Y	3					N
8					N	F 1					N
9					N	2	X	X	X	X	Y
10	X	X	X		Y	T 1					N
A 1				X	N	Me 1			X		N
2	X	X	X		Y	2				X	N
3	X	X	X		Y	3					N
4					N	G 1	X	X	X	X	Y
5	X		X		Y	2					N

KEY

N - Number Numeration
A - Addition + Subtraction
M - Multiplication + Division
F - Fractions

G - Geometry
Me- Measurement
Mo- Money
T - Time

Content Validity Coefficient .50

Judge Reliability .81

Table 6F

RESULTS OF CONTENT VALIDITY ASSESSMENT
SRA TEST AND THE STA1 PROGRAM

Objective	1	2	3	4	Tested	Objective	1	2	3	4	Tested
N 1					N	A 2					N
2					N	3					N
3					N	4					N
4			X		N	5	X				N
5					N	6		X		X	Y
6					N	7			X		N
7					N	Mo 1	X				N
8	X	X	X	X	Y	2		X	X	X	Y
9	X	X	X	X	Y	3					N
10			X		N	4					N
11	X	X		X	Y	G 1					N
12					N	2					N
13					N	3					N
14	X	X	X		Y	Me 1					N
A 1				X	N	2					N

KEY N - Number Numeration
 A - Addition + Subtraction
 M - Multiplication + Division
 F - Fractions

 G - Geometry
 Me- Measurement
 Mo- Money
 T - Time

Content Validity Coefficient .20

Judge Reliability .94

Table 6G

RESULTS OF CONTENT VALIDITY ASSESSMENT
SRA TEST AND THE AW2 PROGRAM

Objective	1	2	3	4	Tested	Objective	1	2	3	4	Tested
N 1					N	M 1			X	X	Y
2	X		X	X	Y	2					N
3					N	3		X			N
4					N	4	X	X		X	Y
5	X	X	X	X	Y	5	X			X	Y
6					N	F 1					N
7	X	X	X		Y	2		X	X		Y
8	X	X	X	X	Y	Me 1					N
9					N	2					N
10					N	3					N
11	X	X	X	X	Y	T 1	X	X		X	Y
12					N	Mo 1					N
13			X		N	2	X	X			Y
A 1	X	X	X	X	Y	3				X	N
2			X		N	G 1					N
3	X		X	X	Y	2					N
4	X	X	X	X	Y	3	X	X		X	Y
5					N	4	X	X		X	Y
6	X	X	X	X	Y	5					N
7					N	6					N
8	X	X	X		Y						

KEY N - Number Numeration
 A - Addition + Subtraction
 M - Multiplication + Division
 F - Fractions

G - Geometry
 Me- Measurement
 Mo- Money
 T - Time

Content Validity Coefficient .44

Judge Reliability .68

Table 6H

RESULTS OF CONTENT VALIDITY ASSESSMENT
SRA TEST AND THE STA2 PROGRAM

Objective	1	2	3	4	Tested	Objective	1	2	3	4	Tested
N 1					N	M 3	X	X	X		Y
2			X		N	4				X	N
3					N	5					N
4	X	X	X	X	Y	6					N
5	X	X	X		Y	7					N
6	X	X	X		Y	Mo 1					N
7				X	N	2	X	X			Y
8					N	3					N
9	X	X	X	X	Y	4					N
10		X		X	Y	G 1					N
A 1				X	N	2					N
2	X		X		Y	3					N
3	X	X	X	X	Y	4					N
4					N	5		X	X	X	Y
5					N	6					N
6	X	X	X	X	Y	7					N
7		X	X		Y	8					N
8	X	X			Y	Me 1	X	X		X	Y
9	X	X	X	X	Y	2					N
10	X	X	X		Y	3					N
M 1					N	4					N
2	X	X	X		Y						

KEY N - Number Numeration
 A - Addition + Subtraction
 M - Multiplication + Division
 F - Fractions

G - Geometry
 Me- Measurement
 Mo- Money
 T - Time

Content Validity Coefficient .40

Judge Reliability .85

results appear in Table 6F.

The SRA Test and the AW2 Program

The content validity coefficient associated with the SRA Test with respect to the AW2 program is .44. Eighteen of the forty-one objectives identified for this program were judged to be tested with an associated judge reliability coefficient of .68.

As shown in Table 6G of five of the thirteen number-numeration objectives and five of the eight addition and subtraction objectives were judged to be tested. Three of the five multiplication objectives and two of the six geometry objectives were also assessed to be tested.

The SRA Test and the STA2 Program

Seventeen of the forty-three objectives identified for the STA2 program were judged to be tested by the SRA Test. The content validity coefficient is thus .40. The judge reliability associated with this assessment of content validity is .85.

A breakdown of the content validity results for the SRA Test on the STA2 program (see Table 6H) reveals that five of the ten number-numeration objectives, seven of the ten addition and subtraction objectives, two of seven multiplication and division objectives and one of eight geometry objectives were judged to be tested.

RESULTS WITH RESPECT TO THE HYPOTHESES

Hypothesis One:

(a) There is no significant difference between the content validity of the Cooperative Test and the SRA Test with reference to

the objectives of the STAl program.

This hypothesis was tested by comparing the proportion of objectives tested by each test. The comparison resulted in the rejection of this null hypothesis. As shown in Table 7 the z-value is 2.921. A value of 1.96 is needed in order that the difference between the proportions be significant at the .05 level. The difference favoured the Cooperative Test.

(b) There is no significant difference between the content validity of the Cooperative Test and the SRA Test with reference to the objectives of the AW1 program.

The comparison by means of a z-test, of the proportion of objectives tested by each test resulted in the null hypothesis being accepted.

(c) There is no significant difference between the content validity of the Cooperative Test and the SRA Test with reference to the objectives of the STA2 program.

The comparison, by means of a z-test, of the proportion of objectives tested by each test resulted in the null hypothesis being accepted.

(d) There is no significant difference between the content validity of the Cooperative Test and the SRA Test with reference to the objectives of the AW2 program.

The comparison, by means of a z-test, of the proportion of objectives tested by each instrument resulted in the null hypothesis being accepted.

Table 7

COMPARISON OF THE CONTENT VALIDITY OF THE COOPERATIVE TEST AND THE SRA TEST
WITH REFERENCE TO THE OBJECTIVES OF

(a) THE STA1 PROGRAM
(b) THE AW1 PROGRAM
(c) THE STA2 PROGRAM
(d) THE AW2 PROGRAM

Program	Proportion of objectives tested by the Cooperative test	Proportion of objectives tested by the SRA test	Z-Value
AW1	16/30	15/30	0.258
STA1	17/30	6/30	2.921 *
AW2	23/41	18/41	1.104
STA2	18/43	17/43	0.219

* significant at the .05 level

Hypothesis Two:

(a) There is no significant difference between the mean achievement scores of pupils in the STA1 program and pupils in the AW1 program as measured by the Cooperative Test.

The results from the analysis of this null hypothesis reveal that it could not be rejected. As indicated in Table 8 at the .05 level there was no significant difference between the two groups on the criteria of mathematics achievement as measured by the Cooperative Test.

(b) There is no significant difference between the mean achievement scores of pupils in the STA1 program and pupils in the AW1 program as measured by the SRA Test.

The results from the analysis of this null hypothesis reveal that it could not be rejected. As indicated in Table 8 at the .05 level there was no significant difference between the two groups on the criteria of mathematics achievement as measured by the SRA Test.

Hypothesis Three:

(a) There is no significant difference between the mean achievement scores of pupils in the STA2 program and pupils in the AW2 program as measured by the Cooperative Test.

The results from the analysis of this null hypothesis reveal that it must be rejected. As indicated in Table 8, at the .05 level there was a significant difference between the two groups on the criteria of mathematics achievement as measured by the Cooperative Test. The difference was in favour of pupils in the AW2 program.

Table 8

ANALYSIS OF COVARIANCE ON THE CRITERION OF
 MATHEMATICS ACHIEVEMENT OF PUPILS IN THE TWO GRADES
 OF THE AW PROGRAM AND THE STA PROGRAM WITH I.Q. AS A COVARIATE

Source	Mean Score AW	STA	Adj. SS	DF	Adj. MS	Adj. P	P
AW1 and STA1 on the Cooperative Test	36.80	35.30	34.80	1	34.80	1.00	0.322
AW1 and STA1 on the SRA Test	16.70	14.70	60.72	1	60.72	3.35	0.073
AW2 and STA2 on the Cooperative Test	47.03	42.70	286.69	1	286.69	11.57	0.001 *
AW2 and STA2 on the SRA Test	28.70	23.53	411.48	1	411.48	12.05	0.001 *

* significant beyond the .05 level

(b) There is no significant difference between the mean achievement scores of pupils in the STA2 program and pupils in the AW2 program as measured by the SRA Test.

The results from the analysis of this null hypothesis reveal that it must be rejected. As indicated in Table 8, at the .05 level there is a significant difference between the two groups on the criteria of mathematics achievement as measured by the SRA Test. The difference was in favour of the AW2 group.

Hypothesis Four:

In measuring achievement, there is no significant difference between the mean score obtained on the Cooperative Test and the mean score obtained on the SRA Test for pupils on (a) the STA1 program (b) the AW1 program (c) the STA2 program (d) the AW2 program.

This hypothesis was tested by converting each child's score to a percent and comparing the mean percents of each test. The comparison was made by means of a t-test. The results are summarized in Table 9.

The analysis of hypothesis 4 (a) resulted in a t-test of 14.624. This caused the null hypothesis to be rejected. The mean score attained by pupils in the STA1 program on the Cooperative Test was significantly superior to that obtained on the SRA Test. This result is significant at the .01 level.

The null hypothesis 4 (b) was also rejected at the .01 level. The t-value obtained was 9.355. Pupils in the AW1 program scored significantly higher on the Cooperative Test than on the SRA Test.

Table 9

COMPARISON OF THE MEAN SCORES ATTAINED ON
THE COOPERATIVE TEST AND THE SRA TEST BY PUPILS IN
(a) THE STA1 PROGRAM
(b) THE AW1 PROGRAM
(c) THE STA2 PROGRAM
(d) THE AW2 PROGRAM

Program	Mean Percent Score on the Cooperative Test	Mean Percent Score on the SRA Test	t-Value	Probability
STA1	64.17	35.00	14.624	0.000 *
AW1	66.91	39.76	9.335	0.000 *
STA2	77.63	56.03	9.757	0.000 *
AW2	85.48	68.33	8.848	0.000 *

* significant at the .05 level

A t-value of 9.757 resulted in the rejection of the null hypothesis 4 (c). Pupils in the STA2 program had significantly higher achievement scores on the Cooperative Test. This result is significant at the .01 level.

The analysis of hypothesis 4 (D) resulted in a t-value of 8.848. This caused the null hypothesis to be rejected. Pupils in the AW2 program scored significantly higher on the Cooperative Test. The result is significant at the .01 level.

Hypothesis Five:

(a) There is no significant difference between the mean achievement scores of pupils in the high ability group of the STA1 program and pupils in the high ability group of the AW1 program as measured by firstly the Cooperative Test and secondly the SRA Test.

The testing of this hypothesis by means of analysis of covariance with IQ as covariate resulted in it being accepted in both cases. As shown in Table 10 there is no significant difference at the .05 level between the mean achievement scores of these two groups on either of the two tests.

(b) There is no significant difference between the mean achievement scores of pupils in the average ability group of the STA1 program and pupils in the average ability group of the AW1 program as measured by firstly the Cooperative Test and secondly the SRA Test.

As illustrated in Table 10 the testing of this null hypothesis by means of analysis of covariance with I.Q. as covariate resulted in it being accepted in both cases at the .05 level.

Table 10

ANALYSIS OF COVARIANCE WITH I.Q. AS THE COVARIATE ON THE CRITERION OF MATHEMATICS ACHIEVEMENT
FOR THE HIGH, AVERAGE AND LOW ABILITY GROUPS (GRADE ONE)

Source	Mean Score AW	STA	SS	DF	MS	Adj. F	P
HAW1 compared to HSTAL on the Cooperative Test	40.70	39.70	0.020	1	0.020	0.0008	0.978
HAW1 compared to HSTAL on the SRA Test	23.10	17.50	44.13	1	44.13	2.27	0.150
AAW1 compared to ASTAL on the Cooperative Test	38.10	35.10	31.10	1	31.10	1.97	0.178
AAW1 compared to ASTAL on the SRA Test	13.60	14.50	0.781	1	0.781	0.051	0.825
LAW1 compared to LSTAL on the Cooperative Test	31.60	31.10	1.66	1	1.66	0.020	0.889
LAW1 compared to LSTAL on the SRA Test	13.40	12.10	8.53	1	8.53	0.54	0.471

(c) There is no significant difference between the mean achievement scores of pupils in the low ability group of the STA1 program and pupils in the low ability group of the AW1 program as measured firstly by the Cooperative Test and secondly by the SRA Test.

Testing this hypothesis by means of analysis of covariance with IQ as covariate resulted in it being accepted in both cases at the .05 level. This is shown in Table 10.

Hypothesis Six:

(a) There is no significant difference between the mean achievement scores of pupils in the high ability group of the STA2 and pupils in the high ability group of the AW2 program as measured firstly by the Cooperative Test and secondly by the SRA Test.

Results of analysis of covariance with IQ as the covariate, illustrated in Table 11, indicate that this null hypothesis should be accepted, at the .05 level, in both cases.

(b) There is no significant difference between the mean achievement scores of pupils in the average ability group of the STA2 program and pupils in the average ability group of the AW2 program as measured by firstly the Cooperative Test and secondly the SRA Test.

The testing of this hypothesis by means of analysis of covariance indicates that it should be accepted at the .05 level in the first case and rejected at the .05 level in the second case. There is a significant difference in the mean achievement scores of these two groups, in favour of the AW2 average group, on the SRA Test. There is no difference in these scores on the Cooperative Test.

Table 11

ANALYSIS OF COVARIANCE WITH I.Q. AS THE COVARIATE ON THE CRITERION OF MATHEMATICS ACHIEVEMENT
FOR THE HIGH, AVERAGE AND LOW ABILITY GROUPS (GRADE TWO)

Source	Mean Score AW2	STA2	Adj. SS	DF	Adj. MS	Adj. F	P
HAW2 compared to HSTA2 on the Cooperative Test	51.50	50.40	12.72	1	12.72	1.94	0.092
HAW2 compared to HSTA2 on the SRA Test	35.20	31.90	59.76	1	59.76	3.19	0.092
AAW2 compared to ASTA2 on the Cooperative Test	48.80	45.10	69.44	1	69.44	2.73	0.117
AAW2 compared to ASTA2 on the SRA Test	29.30	23.90	200.51	1	200.51	5.02	0.039 *
LAW2 compared to LSTA2 on the Cooperative Test	40.80	32.60	209.36	1	209.36	4.70	0.045 *
LAW2 compared to LSTA2 on the SRA Test	21.60	14.80	106.53	1	106.53	2.28	0.150

* significant at the .05 level

(c) There is no significant difference between the mean achievement scores of pupils in the low ability group of the STA2 program and pupils in the AW2 program as measured by firstly the Cooperative Test and secondly the SRA Test.

The results of the analysis of covariance indicate that this hypothesis should be rejected in the first case and accepted in the second. There is no significant difference in the scores of these two groups on the SRA Test but the low AW2 group did significantly better than the low STA group on the Cooperative Test. The level of significance is the .05 level.

Hypothesis Seven

There is no significant difference between the mean achievement scores of pupils in the STA1 program and the AW1 program on the basic facts subtest of the Cooperative Test (b) the SRA Test.

The testing of this null hypothesis by means of a t-test resulted in the acceptance of both hypothesis 7 (a) and 7 (b). The mean achievement scores of pupils in the STA1 program and the AW1 program on the basic facts subtests of both Cooperative and the SRA Tests could not be considered significantly different at the .05 level. The results are summarized in Table 12.

Hypothesis Eight

There is no significant difference between the mean achievement scores of pupils in the STA1 program and the AW1 program on the number-numeration subtest of (a) the Cooperative Test (b) the SRA Test.

Table 12

COMPARISON OF RESULTS ATTAINED BY GRADE ONE PUPILS
IN THE STA PROGRAM AND THE AW PROGRAM ON TWO SUB-TESTS
OF (a) THE COOPERATIVE TEST
(b) THE SRA TEST

Source	STA Mean Score	AW Mean Score	t - Value	Probability
Grade One on the basic facts sub-test of the Cooperative Test	9.10	10.50	1.991	0.051
Grade One on the number- numeration sub-test of the Cooperative Test	14.57	13.93	-0.900	0.372
Grade One on the basic facts sub-test of the SRA Test	7.03	8.30	1.594	0.116
Grade One on the number- numeration sub-test of the SRA Test	2.57	3.07	1.147	0.256

The testing of this null hypothesis by means of a t-test resulted in the acceptance of both hypothesis 7 (a) and 7 (b). The achievement scores of pupils in the STA1 program and the AW1 program on the number-numeration subtests of both the Cooperative and the SRA Tests could not be considered significantly different at the .05 level. The results are summarized in Table 12.

Hypothesis Nine:

There is no significant difference between the mean achievement scores of pupils in the STA2 program and the AW2 program on the basic facts subtest of (a) the Cooperative Test (b) the SRA Test.

The test of this null hypothesis by means of a t-test resulted in the rejection of both hypothesis 9(a) and 9(b). As shown in Table 13 the t-value for hypothesis 9(a) is 2.492 which is significant at the .05 level. The t-value for hypothesis 9(b) is 2.966 which is significant beyond the .01 level. Thus the mean achievement scores of pupils in the AW2 program on the basic facts subtest of both the Cooperative and SRA Tests could be considered significantly better than those of pupils in the STA2 program.

Hypothesis Ten:

There is no significant difference between the mean achievement scores of pupils in the STA2 program and the AW2 program on the number-numeration subtest of (a) the Cooperative Test (b) the SRA Test.

The testing of this null hypothesis by means of a t-test resulted in the acceptance of both hypothesis 10(a) and 10(b). The scores of pupils in the STA2 program and the AW2 program on the number-numeration subtests of both the Cooperative Test and the SRA Test could not be considered significantly different at the .05

Table 13

COMPARISON OF RESULTS ATTAINED BY GRADE TWO PUPILS
IN THE STA PROGRAM AND THE AW PROGRAM ON TWO SUB-TESTS
OF (a) THE COOPERATIVE TEST
(b) THE SRA TEST

Source	STA Mean Score	AW Mean Score	t - Value	Probability
Grade Two on the basic facts sub-test of the Cooperative Test	12.03	13.67	2.492	0.016 *
Grade Two on the number- numeration sub-test of the Cooperative Test	15.83	16.77	1.477	0.145
Grade Two on the basic facts sub-test of the SRA Test	11.63	15.43	2.966	0.004 *
Grade Two on the number- numeration sub-test of the SRA Test	4.20	5.30	1.889	0.064

* significant at the .05 level

Table 14

RESULTS OF THE ANALYSIS OF COVARIANCE USING I.Q. AS COVARIATE
 COMPARING THE PUPILS ANSWERING THE SRA TEST IN MACHINE-MARKABLE FORMAT
 AND PUPILS ANSWERING IN HAND-SCORING FORMAT AT BOTH GRADE LEVELS ON THE CRITERION
 OF MATHEMATICS ACHIEVEMENT

Source	Mean Score		DF	Adj. SS	Adj. MS	Adj. F	P
	Machine	Hand					
Grade One	16.1	14.9	1	5.86	5.86	.215	.645
Grade Two	25.5	26.1	1	4.32	4.32	.080	.779

level. These results are presented in Table 13.

Hypothesis Eleven:

There is no significant difference between the mean achievement scores of pupils writing the machine-markable test and those writing the hand scored test in (a) grade one (b) grade two.

This null hypothesis was accepted in both cases. Results from the analysis of covariance, using IQ as covariate, showed no significant difference between the achievement scores of pupils writing the machine markable SRA Test and those writing the test in hand scoring format at both grade levels. The results are presented in Table 14.

Hypothesis Twelve:

There is no significant difference in the number of items judged not suitable on the Cooperative Test compared to the number judged not suitable on the SRA Test for pupils in (a) the STA1 program (b) the AW1 program (c) the STA2 program (d) the AW2 program.

This hypothesis was tested by comparing the proportion of items judged not suitable on each test. The comparison was made by using a z-test. The results are summarized in Table 15.

The comparison of the proportions of items judged not suitable for pupils in the STA1 program resulted in a z-value of 1.641 and the acceptance of the null hypothesis 12 (a).

The comparison of the proportions of items judged not suitable for pupils in the AW1 program resulted in a z-value of 0.082 and the acceptance of the null hypothesis 12(b).

Table 15

COMPARISON OF THE NUMBER OF ITEMS JUDGED "NOT SUITABLE" FOR PUPILS IN

(a) THE STA1 PROGRAM
 (b) THE AW1 PROGRAM
 (c) THE STA2 PROGRAM
 (d) THE AW2 PROGRAM

Program	Proportion of items judged not suitable on the Cooperative Test	Proportion of items judged not suitable on the SRA Test	Z Value
STA1	16/55	19/42	1.641
AW1	14/55	11/42	0.082
STA2	9/55	3/42	1.367
AW2	17/55	4/42	2.434 *

* significant at the .05 level

The null hypothesis 12(c) was accepted at the .05 level. The z-value, obtained by comparing the items on each test judged not suitable for pupils in the STA2 program, was 1.367.

The comparison of the proportions of items judged not suitable for pupils in the AW2 program resulted in a z-value of 2.534 and the acceptance of the null hypothesis 12(d).

RESULTS OF ADDITIONAL ANALYSIS

The two tests were compared for each of the four mathematics programs on the basis of the following criteria; content validity, valid items, unsuitable items, mean score for each of the three ability levels, subtest scores. The results of these comparisons are summarized below and presented in Tables 16 (A), (B), (C) and (D). The results obtained from testing hypothesis four are included in these tables, also, in order that a more complete picture may be presented.

The scores obtained by pupils in each of the three ability groups of the AW1 program on the Cooperative Test were significantly higher than their scores on the SRA Test. Pupils in this program also did significantly better on each of the two major subtests of the Cooperative Test than on the corresponding subtests of the SRA Test. Differences in content validity, valid items and unsuitable items were not significant although in each case, the Cooperative Test was favoured.

TABLE 16A

SUMMARY OF COMPARISONS MADE BETWEEN THE CO-OPERATIVE AND SRA
TESTS WITH RESPECT TO THE AWI PROGRAM

Comparison	Co-operative Test	SRA Test
Content Validity	16/30 (.53)	15/30 (.50)
Difficulty	.67	.41
Reliability	.84	.82
Valid Items	36/55 (65%)	18/42 (43%)
Unsuitable Items	14/55 (20%)	11/42 (24%)
MEAN SCORE		
(1) Whole Groups	36.80 (67%) *	16.70 (40%)
(2) High Group	40.70 (74%) *	23.10 (55%)
(3) Average Group	38.10 (69%) *	13.60 (32%)
(4) Low Group	31.60 (57%) *	13.40 (32%)
SUBTEST SCORES		
(1) Basic Facts	10.50 (62%) *	8.30 (38%)
(2) Number-Numeration	13.93 (73%) *	2.57 (38%)

* The difference is significant at the .05 level.

TABLE 16B

SUMMARY OF COMPARISONS MADE BETWEEN THE CO-OPERATIVE AND SRA
TESTS WITH RESPECT TO THE STA 1 PROGRAM

Comparison	Co-operative Test	SRA Test
Content Validity	17/30 (.57) *	6/30 (.20)
Difficulty	.64	.37
Reliability	.76	.61
Valid Items	29/55 (53%) *	7/42 (17%)
Unsuitable Items	16/55 (29%)	19/42 (45%)
MEAN SCORE		
(1) Whole Groups	35.30 (64%) *	14.70 (35%)
(2) High Group	39.70 (72%) *	17.50 (42%)
(3) Average Group	35.10 (64%) *	14.50 (35%)
(4) Low Group	31.10 (57%) *	12.10 (29%)
SUBTEST SCORES		
(1) Basic Facts	9.10 (54%) *	7.03 (32%)
(2) Number-Numeration	14.57 (77%)	3.07 (32%)

TABLE 16C

SUMMARY OF COMPARISONS MADE BETWEEN THE CO-OPERATIVE AND SRA
TESTS WITH RESPECT TO THE AW2 MATHEMATICS PROGRAM

Comparison	Co-operative Test	SRA Test
Content Validity	23/41 (.56)	18/41 (.44)
Difficulty	.86	.67
Reliability	.80	.90
Valid Items	45/55 (82%)	33/42 (79%)
Unsuitable Items	17/55 (21%) *	4/42 (10%)
MEAN SCORE		
(1) Whole Group	47.03 (83%) *	28.70 (66%)
(2) High Group	51.50 (94%) *	35.20 (84%)
(3) Average Group	48.80 (89%) *	29.30 (70%)
(4) Low Group	40.80 (74%) *	21.60 (51%)
SUBTEST SCORES		
(1) Basic Facts	13.67 (77%) *	15.43 (68%)
(2) Number Numeration	16.77 (85%) *	5.30 (66%)

* The difference is significant at the .05 level.

TABLE 16D

SUMMARY OF COMPARISONS MADE BETWEEN THE CO-OPERATIVE AND SRA
TESTS WITH RESPECT TO THE STA 2 PROGRAM

Comparison	Co-operative Test	SRA Test
Content Validity	18/43 (.45)	17/43 (.40)
Difficulty	.78	.58
Reliability	.93	.93
Valid Items	35/55 (64%)	23/42 (55%)
Unsuitable Items	9/55 (18%)	3/42 (7%)
MEAN SCORE		
(1) Whole Groups	42.70 (78%) *	23.53 (56%)
(2) High Group	50.40 (92%) *	31.90 (76%)
(3) Average Group	45.10 (82%) *	23.90 (57%)
(4) Low Group	32.60 (59%) *	14.80 (35%)
SUBTEST SCORES		
(1) Basic Facts	12.03 (71%) *	11.63 (53%)
(2) Number Numeration	15.83 (83%) *	4.20 (53%)

* The difference is significant at the .05 level.

The content validity of, and the number of valid items on the Cooperative Test were significantly higher than for the SRA Test with respect to the STA1 program. Pupils in this program also scored significantly higher on the Cooperative Test than on the SRA Test in each of the five categories mentioned above. The number of unsuitable items on each test was not significantly different although there were more on the SRA Test.

Differences between the two tests on the criteria of content validity, valid items, unsuitable items were not significant with respect to the AW2 program. Differences in scores obtained by each of the three ability groups were significant in favour of the Cooperative Test. Pupils in the AW2 program also scored significantly higher on both the number-numeration and basic facts subtests of the Cooperative Test than on these subtests of the SRA Test.

The high, average and low ability groups of the STA2 program scored significantly higher on the Cooperative Test than on the SRA Test. Pupils in this program also had significantly higher scores on the two subtests of the Cooperative Test than those of the SRA Test. The difference in content validity, valid items and unsuitable items were not significant.

SUMMARY OF THE RESULTS

The results of the data analysis may be summarized as follows:

(1) There was no significant difference between the content validity of the Cooperative Test and the SRA Test with respect to the objectives of (a) the AW1 program (b) the AW2 program (c) the STA2

program. The content validity of the Cooperative Test is significantly higher than that of the SRA Test with respect to the objectives of the STA1 program.

(2) There was no significant difference between the mean achievement score of pupils in the AW1 program and the STA1 program as measured either the Cooperative or SRA Tests.

(3) Pupils in the AW2 program scored significantly higher than pupils on the STA2 program on both the Cooperative and SRA Tests.

(4) Achievement scores of pupils in all four mathematics programs were significantly higher on the Cooperative Test than on the SRA Test.

(5) There was no significant difference between the pupils in each of the three ability groups of the AW1 program and the three ability groups of the STA1 program on the criteria of mathematics achievement as measured by both the Cooperative and SRA Tests.

(6) Pupils in the average ability group of the AW2 program scored significantly higher on the SRA Test than pupils in the average ability group of the STA2 program. The low ability group of the AW2 program scored significantly higher on the Cooperative Test than pupils in the low ability groups at the STA2 program. There was no significant difference between scores of the high ability groups of the two programs as measured by either test. Also there was no significant difference between the scores of the average groups of the two programs on the Cooperative Test and the two low groups on the SRA Test.

(6) Pupils in the average ability group of the AW2 program scored significantly higher on the SRA Test than pupils in the average ability group of the STA2 program. The low ability group of the AW2 program scored significantly higher on the Cooperative Test than pupils in the low ability groups at the STA2 program. There was no significant difference between scores of the high ability groups of the two programs as measured by either test. Also there was no significant difference between the scores of the average groups of the two programs on the Cooperative Test and the two low groups on the SRA Test.

(7) There is no significant difference between the scores of pupils on the AW1 program and the STA1 program on both the number-numeration and basic facts subtests of both the Cooperative and SRA Tests.

(8) Pupils in the AW2 program scored significantly higher than pupils in the STA2 program on the basic facts, subtests of both the Cooperative and SRA Tests. There was no significant difference in their scores on the number-numeration subtests of either test.

(9) There is no significant difference in the scores of pupils answering the SRA Test in the machine markable format compared to those answering in the hand scoring format at both grade levels.

(10) The number of unsuitable items on the Cooperative Test was not significantly different than the number judged unsuitable on the SRA Test for each of the AW1, STA1, STA2 programs. The Cooperative Test contained significantly more unsuitable items than the SRA Test with reference to the AW2 program.

(11) Further analysis revealed that for all four programs a comparison of scores on the Cooperative and SRA Tests, whether this comparison be made on the mean score for both tests on the two subtests, pupils in each of the programs scored significantly higher on the Cooperative Test than on the SRA Test. This result was consistent for each of the ability groups.

TABLE 17

SUMMARY OF INFORMATION CONCERNING THE COOPERATIVE TEST

Source of Information	AW1	STA1	AW2	STA2
Content Validity	16/30	17/30	23/41	18/43
Difficulty	.672	.642	.855	.777
Valid Items	36/55*	29/55	45/55*	35/55
No. of Unsuitable Items	14/55	16/55	17/55*	9/55
TEST MEANS				
(1) Whole Group	36.80	35.30	47.03*	42.70
(2) High Achievement Group	40.70	39.70	51.50	50.40
(3) Average Achievement Group	38.10	35.10	48.80	45.10
(4) Low Achievement Group	31.60	31.10	40.80*	32.60
Reliability	.84	.76	.90	.93
Basic Facts	10.50	9.10	13.67	12.03
Number-Numeration	13.93	14.57	16.77	15.83

TABLE 18

SUMMARY OF INFORMATION CONCERNING THE SRA TEST

Source of Information	AW1	STA1	AW2	STA2
Content Validity	15/30*	6/30	18/41	17/40
Difficulty	.414	.366	.665	.581
Valid	18/42*	7/42	33/42*	22/42
Unsuitable Items	11/42	19/42	4/42	3/42
TEST MEANS				
(1) Whole Group	16.70	14.70	28.70*	23.53
(2) High Achievement Group	23.10	17.50	35.20	31.90
(3) Average Achievement Group	13.60	14.50	29.30*	23.90
(4) Low Achievement Group	13.40	12.10	21.60	14.80
Reliability	.82	.61	.90	.93
SUBTEST SCORES				
Basic Facts	8.30	7.03	15.43*	11.63
Number-Numeration	2.57	3.07	5.30	4.20

Chapter 5

SUMMARY, CONCLUSIONS, IMPLICATIONS AND SUGGESTIONS FOR FURTHER RESEARCH

SUMMARY

The purpose of this research was to evaluate two standardized mathematics achievement tests, the Cooperative Primary Test, Mathematics, Form 12A and the SRA Modern Mathematics Understanding Test, Primary Level Form C, in order to determine their suitability for measuring mathematics achievement in two primary mathematics programs in Alberta.

One hundred and twenty pupils participated in the study. These pupils were selected from six schools, three of which followed the mathematics program of the Addison-Wesley text series and three which followed the Seeing Through Arithmetic Series. Thirty pupils were chosen from each of grades one and two of the two mathematics programs. The pupils were chosen by the school personnel on the basis of mathematics ability. They were asked to choose equal numbers of pupils in each of three ability groups at each grade level.

The content validity of the two tests with respect to the two grade levels of each of the mathematics programs was assessed by submitting the tests to the scrutiny of a committee of judges. The judges were asked to match test items to program objectives and thus the number of objectives of each of the four programs tested by each instrument was determined.

The two tests were then administered to each of the one

hundred and twenty pupils by the researcher. The achievement scores so obtained were then compared.

Scores of pupils in the AW program were compared with scores of pupils in the STA program at each grade level. These comparisons were made for scores on the complete test and for scores on the basic facts and number-numeration subtests of each test. The scores of pupils in each of the three ability groups at each grade level of the AW program were compared to the scores of the matching ability groups in the STA program. These comparisons were based on total test scores. The mean scores obtained on each of the two tests were also compared for each of the four mathematics program.

The SRA Test is machine markable. A comparison was made of the scores of pupils answering the test using this format with those answering in hand scoring format.

The numbers of unsuitable items on each test for each of the four mathematics programs was identified and compared.

By means of an item analysis a difficulty index for each item on the two tests was determined with respect to each of the four programs. By averaging the difficulty indexes, a difficulty rating for both tests was determined for each of the four programs.

On the basis of all the data obtained, an evaluation of each of the tests was made as to their suitability for use as instruments for measuring achievement in two primary mathematics programs in Alberta.

CONCLUSIONS

CONCLUSIONS WITH RESPECT TO THE MAJOR HYPOTHESES

Hypothesis One:

On the basis of content validity the Cooperative Test could be used as an instrument for measuring mathematics achievement at the primary level without a bias in favour of either program.

The SRA Test could be used at the grade two level as there is no difference in the content validity of this test with respect to either program. However, it is not suitable at the grade one level as there is a significant bias in favour of the AW1 program.

Hypothesis Two:

Pupils in the AW1 and STAl mathematics programs score equally well on both the Co-operative and SRA Tests.

The content validity coefficient of the Cooperative Test with respect to the AW1 program is .53, the test reliability .85, the difficulty index .67 and the number of valid items thirty-six. For the STAl program the Co-operative Test had a content validity of .57, reliability .76 difficulty index .64 and twenty-nine valid items. The content validity of this test with respect to the two programs may be considered to be equal as can the difficulty indices. However, there is a significant difference with respect to the number of valid items in favour of the AW1 program. Thus one could expect the AW1 pupils to obtain higher scores than the STAl pupils. In fact, the low reliability of this test with respect to the STAl sample as compared to that reported in the test handbook may suggest guessing on the

part of these pupils. This may account for the fact that the mean score is considerably higher than the number of valid items.

The suspicion that the STA1 pupils guessed at the answers to a number of items on these tests is given added weight when one looks at the results from the SRA Test. Here there were significant differences in the content validity (.50 to .20) and the number of valid items (eighteen to seven) in favour of the AW1 program. The test appeared to be less difficult for the AW1 pupils also as the difficulty index for the AW1 pupils was .41 as compared to .37 for the STA1 pupils. The test reliability of the SRA Test for the STA1 pupils though is .61 which is quite low. Thus the fact that there was no significant difference in the mean scores of the AW1 and the STA1 pupils may be explained by the possibility of guessing on the part of the STA1 pupils.

Hypothesis Three:

Pupils in the AW2 mathematics program score significantly higher on both the Cooperative and SRA Tests than do pupils in the STA2 program.

The fact that pupils in the AW2 program scored significantly higher on both the Cooperative and SRA Tests is likely due to the significant differences in the number of valid items in favour of the AW2 program on both tests and the differences in difficulty indexes of both tests in favour of the AW2 program.

The number of valid items on the Cooperative Test for the AW2 program was forty-five as compared to thirty-five for the STA2 program. The difficulty index for the AW2 program was .86 and for the STA2 program .78.

The number of valid items on the SRA Test favoured the AW2 program thirty-three to twenty-two and the difficulty index .67 to .58.

Hypothesis Four:

Pupils in all four mathematics programs score significantly higher on the Cooperative Test than on the SRA Test.

The difficulty indexes of the Cooperative Test for the AW1, AW2 and STA2 mathematics programs were .67, .86 and .78 respectively. The corresponding difficulty indexes of the SRA Test were .41, .66 and .58 respectively. The fact that scores for pupils in all three programs were significantly higher on the Cooperative Test results from the differences in difficulty.

In the case of the STA1 program, not only was there a difference in difficulty index (.64 to .37) in favour of the Cooperative Test, but differences in content validity (.57 to .20) and number of valid items (twenty-nine to seven) were statistically significant in favour of the Cooperative Test. Thus the score of the STA1 pupils on the Cooperative Test would be expected to be significantly higher than their score on the SRA Test.

Hypothesis Five:

Pupils in the high ability group of the AW1 program and the high ability group of the STA1 program score equally well on firstly the Cooperative Test and secondly the SRA Test. Similar conclusions may be drawn regarding pupils in the average and low ability groups in both programs.

The results from this hypothesis may be interpreted in the same manner as for Hypothesis Two. In view of the content validity, difficulty and number of valid items one might have expected the AW1 pupils to score significantly higher than the STA1 pupils on both tests. The fact that they did not may be an indication that pupils in the STA1 program guessed at answers with a measure of success.

Hypothesis Six:

Pupils in the low ability group of the AW2 program score significantly higher than pupils in the low ability group of the STA2 program on the Cooperative Test. These two groups score equally well on the SRA Test.

Pupils in the average ability group of the AW2 program score significantly higher on the SRA Test than do pupils in the average group of the STA2 program. The two groups score equally well on the Cooperative Test.

Pupils in the high ability groups of both programs score equally well on both the Cooperative and SRA Tests.

Since there is a significant difference in the number of valid items on the Cooperative Test in favour of the AW2 pupils one might expect that the AW2 pupils would do significantly better on the Cooperative Test at all three ability levels than pupils in the STA2 program. That only the low ability group of the AW2 program scored significantly higher is probably due to the relative ease of this test for both grade two samples. Because of this lack of difficulty pupils in the average and high groups of the STA2 program were not hindered by the difference in number of valid items. The low ability group of the STA2 program however was not able to overcome this

handicap.

The superiority of the AW2 group in achievement on the SRA Test is a direct result of the difference in difficulty and number of valid items. The high ability group of the STA2 sample were partially able to overcome this handicap and the difference in scores here is not statistically significant. However, the difference between the two average ability groups is significant. The difference between the two low ability groups while not statistically significant is greater than the difference between the two average groups.

Hypotheses Seven and Eight:

Pupils in the AW1 and STA1 mathematics programs score equally well on the number-numeration and basic facts subtests of both the Cooperative and SRA Tests.

The results of these two hypotheses in which the AW1 pupils and the STA1 pupils are compared on the basic facts and number-numeration subtests of the Cooperative Test are an accurate reflection of the content validity, difficulty and number of valid items (Appendix E) of these tests with respect to the two mathematics programs. There is no significant difference in the content validity and number of valid items of both these subtests in favour of either program. The scores obtained by pupils in the two programs are likewise not significantly different.

A comparison of the content validity of and the number of valid items on the basic facts subtest of the SRA Test with respect to the AW1 and STA1 mathematics programs reveals a significant difference in favour of the AW1 program in both content validity and

number of valid items. The results however, do not reflect these differences. There is no significant difference between the scores of the AW1 pupils and the STA1 pupils on this subtest. This may be attributed to guessing on the part of the STA1 pupils as indicated by the overall low test reliability.

It is interesting to note that although one would not expect it in terms of content validity, number of valid items and difficulty, pupils in the STA1 program scored higher on the number-numeration subtests of both the Cooperative and SRA Tests. Neither difference however was significant.

Hypotheses Nine and Ten:

Pupils in the AW2 program scored significantly higher on the basic facts subtest of both the Cooperative and SRA Tests. They scored equally well on the number-numeration subtest of both tests.

A comparison of the content validity of and the number of valid items on the two subtests of the Cooperative Test with respect to the AW2 and the STA2 mathematics programs revealed no significant differences. The difference in the difficulty indexes of the two subtests with respect to the two programs is not great. However, any differences that do exist favour the AW2 program. A combination of these differences may have resulted in the significant difference in the scores on the basic facts subtest of the Co-operative Test.

A similar argument may be presented to explain the significant difference in the achievement scores on the basic facts subtest of the SRA Test in favour of the AW2 program.

It is again interesting to note that in this case, although the pupils in the AW2 program scored significantly higher on the basic facts subtests of both tests, there were no significant differences in the scores on the number-numeration subtests.

Hypothesis Eleven:

Pupils in both grade one and grade two who answered the test using the hand scored format scored as well as pupils answering in the machine-scorable format of the SRA Test.

Hypothesis Twelve:

The Cooperative Test does not contain more items judged unsuitable than does the SRA Test for pupils in the AW1, STA1 and STA2 programs. For pupils in the AW2 program the Co-operative Test contains more unsuitable items than does the SRA Test.

With regard to evaluating these two instruments for use in the schools in Alberta the following interpretation may be given to the results of hypothesis twelve.

The Cooperative Test may not be suitable as a post test in grade two as it is not sufficiently difficult. The mean difficulty index for the seventeen items judged unsuitable for the AW2 program was .93; for the nine items judged unsuitable for the STA program .87 and for the entire grade two sample the eight unsuitable items had a mean difficulty index of .86.

The SRA Test would seem too difficult for use with grade one pupils. The eleven unsuitable items for the AW1 program had a mean difficulty index of .33; for the STA1 program the nineteen unsuitable items had a mean difficulty index of .29; and the eleven unsuitable items on the SRA Test for the entire grade one sample had

a mean difficulty index of .26.

CONCLUSIONS WITH RESPECT TO ADDITIONAL ANALYSIS

Pupils in grades one and two in both mathematics programs used in this study scored higher on the Cooperative Test than on the SRA Test.

These results are an accurate reflection of the content validity of, the number of valid items on, and the difficulty of the Cooperative and SRA Test with respect to each of the four programs. In each of the four cases the content validity and number of valid items favour the Cooperative Test although in only one case (STAl) are the differences significant. Also, for each of the four groups the Cooperative Test had a higher difficulty index than the SRA Test. In view of these facts one would expect scores obtained on the Cooperative Test to be higher than those obtained on the SRA Test.

IMPLICATIONS

The Cooperative Primary Test, Mathematics, Form 12A

This is a well constructed test which meets the requirements of all the criteria for selection of a standardized test, except perhaps in the area of content validity. A decision concerning the content validity would have to be made by prospective users of the test based on the purpose to which the test is to be put.

This test would be most useful as a post test in grade one or a pre-test in grade two. The difficulty index of this test associated with the grade two sample used in the study indicates

that the score would be such that the distribution of scores would be negatively skewed thus presenting an incomplete picture of the pupils achievement.

This test has the advantage of an alternate form, thus one form could be used as a post-test in grade one and the other a pre-test in grade two.

The SRA Modern Mathematics Understanding Test, Primary, Form C

This test would be most useful as a post test in grade two. Although designed for use both in grades one and two, it seems too difficult to be a valuable instrument for use in grade one or as a pre-test in grade two, although alternate forms are available. It would seem more appropriate to use this test as a pre-test in grade three.

The same comments concerning content validity may be made concerning the SRA Test as were made in the section on the Cooperative Test. A decision concerning the content validity would be a value judgement on the part of a prospective user.

With regard to the criteria stated in Chapter Three, the SRA Test fails to meet them in two specific areas. In the first case, there is no guide book available to explain the construction or the norming of the test or indeed what the norms are. Secondly the price may be too high as the marking service supplied by SRA must be purchased when one uses the Modern Mathematics Understanding Test, Primary Form.

A comment concerning the actual content of the test seems appropriate. Twenty-two of the forty-two items on this test were identified as the basic facts subtest of this test. Eight items were identified as the number-numeration subtest. Thus the stress is in favour of the computational aspects of mathematics.

The claim by SRA that children in grades one and two are able to handle the machine-marking format of this test seems to be substantiated by the results of this study.

SUGGESTIONS FOR FURTHER RESEARCH

This study attempted to evaluate two primary mathematics achievement tests in terms of their usefulness as instruments in the mathematics program currently in use in the Province of Alberta. Similar studies could be undertaken to evaluate tests at the higher grade levels.

In this regard, the content validity of the Cooperative Primary Test, Mathematics, Form 23A was assessed with respect to the objectives of both the AW2 and STA2 mathematics programs (Appendix F) the number of valid items on this test for each of the two programs was also determined.

One interesting side result of this study was the performance of pupils in the STA program on the number-numeration subtests. Further investigation may reveal whether or not this pattern is persistent. If so, perhaps a study could be initiated to evaluate the mathematics programs now in use in the primary grades in Alberta in order to determine their relative usefulness as aids in children's learnings of the meanings in mathematics.

In the opinion of the researcher the content validity of both the Co-operative and SRA Tests is insufficient to make them useful as instruments for measuring mathematics achievement in the STA and AW programs. If such tests are desirable then their construction should be undertaken using the objectives suggested in the Department of Education Curriculum Guide as a basis for constructing items. It is further suggested that the pattern used in the construction of the Cooperative Test be followed, that is one test for each grade level and alternate forms of each test.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Ahmann, J. Stanley; Glock, Marvin D; and Wardeberg, Helen L.
Evaluating Elementary School Pupils. Allyn and Bacon Inc., 1960.
- Ashlock, Robert B. "A Test of Understandings for the Primary Grades,"
Arithmetic Teacher 15:438-41, May, 1968.
- Ashlock, Robert B. and Welch, Ronald C. "A Test of Understandings of
Selected Properties of a Number System: Primary Form,"
Bulletin of the School of Education, Indiana Univesity 42,
March, 1966.
- Begle, Edward G. and Wilson, James W. "Evaluation of Mathematics
Programs." The Sixty-Ninth Yearbook of the National Society
for the Study of Education. NSSE, Chicago, 1970.
- Biggs, E. E., "Mathematics in Primary Schools," in New Approaches
to Mathematics Teaching ed. F. W. Land. MacMillan Co. Ltd.,
London, 1963.
- Brownell, William A. "Psychological Consideration for the Learning
and Teaching of Arithmetic," Thirty-fifth Yearbook National
Council of Teachers of Mathematics. Teachers' College, Columbia
University, New York, 1935.
- Brownell, William A. and Moser, Harold E., Meaningful vs. Mechanical
Learning: A Study in Grade III Subtraction, Duke University
Press, Durham, N. C. 1949.
- Brueckner, Leo J. "Evaluation in Arithmetic" Education 79. 291-294,
January, 1959.
- DeCecco, John P., The Psychology of Learning and Instruction:
Educational Psychology. Prentice-Hall, Englewood Cliffs,
New Jersey, 1968.
- DeVault, M. Vere and Kriewall, Thomas E. Perspectives in Elementary
School Mathematics. Charles E. Merrit Publishing Company,
Columbus Ohio, 1969.
- DeVault, M. Vere; Fennema, Elizabeth; Neufeld, K. Allen; and Smith,
Lewis B., Wisconsin Contemporary Test of Elementary Mathematics
Personnel Press, Inc. Princeton, New Jersey, 1967.
- Douglas, Harl R. and Spitzer, Herbert F. "The Importance of Teaching
for Understanding," The Forty-fifth Yearbook of the National
Society for the Study of Education Part 1. NSSE, Chicago, 1946.

- Dutton, Wibur H. Evaluating Pupils' Understanding of Arithmetic. Prentice-Hall, Englewood Cliffs, New Jersey, 1964.
- Edwards, Charles W., Jr. and Welch, Ronald C. "A Test of Arithmetic Principles Elementary Form, Bulletin of the school of Education, Indiana University 41, September, 1965.
- Epstein, Marion G., "Testing in Mathematics: Why? What? How?" The Arithmetic Teacher 15:311 - 319, April, 1968.
- Ferguson, George A. Statistical Analysis in Psychology and Education. McGraw Hill Inc. 1949.
- Feifel, Herman, and Lorge, Irving, "Qualitative Differences in Vocabulary Responses of Children," Journal of Education Psychology 41: 1-18, January, 1950.
- Flournoy, Francis, "The Development of Arithmetic Understanding Tests for Primary and Intermediate Levels," Journal of Education Research 62: 73-77, October, 1968.
- Garret, Henry E. Testing for Teachers. American Book Company, New York, 1959.
- Gibb, E. Glenadine, "Some Approaches to Mathematics Concepts," NEA Journal 48: 65-66, November, 1959.
- Glennon, Vincent J. "A Study of the Growth and Mastery of Certain Basic Mathematical Understandings on Seven Educational Levels," unpublished doctoral dissertation, Harvard University, 1948.
- Gray, Roland F. "An Approach to Evaluating Arithmetic Understandings," Arithmetic Teacher 13: 187-192, March 1966.
- Greene, Harry A. and Jorgensen, Albert N. The Use and Interpretation of Educational Tests. Longmans, Greene and Co., New York-London - Toronto, 1929.
- Greene, Harry A., Jorgensen, Albert N. and Gerberich, Raymond J. Measurement and Evaluation in the Elementary School. Longman, Green and Co., New York - London - Toronto, 1953.
- Gronlund, Norman E., Constructing Achievement Tests, Prentice-Hall, Englewood Cliffs, New Jersey, 1968.
- Hildreth, Gertrude, "Principles of Learning Applied to Arithmetic," The Arithmetic Teacher 1: 1-5 October, 1959.
- Hoepfner, Ralph (ed.) C.S.E. Elementary School Test Evaluations. Center for the Study of Evaluation, UCLA Graduate School of Education. Los Angeles, California, 1970. (pages 9-16)

- Johnson, Donovan A. "Introduction." Twenty-sixth Yearbook of the National Council of Teachers of Mathematics, The N. C. T. M., Washington, D. C. 1961.
- Johnson, Donovan A., and Trimble, H. C. "Evaluation of Mathematical Meanings and Understandings." Twenty-second Yearbook of the National Council of Teachers of Mathematics, The N.C.T.M., Washington, D.C., 1954.
- Judd, Charles Hubbard, "The Relation of Special Training to General Intelligence." Educational Review 36: 28-42, December 1908.
- Koenker, Robert H. "Measuring the Meanings of Arithmetic," Arithmetic Teacher 7:93-96, February 1960.
- Krich, Percy, "Meaningful vs. Mechanical Method, Teaching Division of Fractions by Fractions," School of Science and Mathematics 64: 697 -708. November, 1964.
- Lindvall, C. M. Measuring Pupil Achievement and Aptitude. Harcourt, Brace and World. New York, 1967.
- Madaus, George F. "Evaluation of a Mathematics Program," Arithmetic Teacher 8: 418-21, December, 1961.
- McSwain, E. T. "Discovering Meanings in Arithmetic" Childhood Education 26: 267-271. February 1950.
- Mehrens, William A. and Lehmann, Irvin J. Holt, Standardized Tests in Education, Rhinehart, Winston, New York, 1969.
- Merwin, Jack C. "Constructing Achievement Tests and Interpreting Scores." Evaluation in Mathematics, Twenty-Sixth Yearbook. National Council of Teachers of Mathematics. Washington, D.C. 1961.
- Miller, G. H., "How Effective Is the Meaning Method?" The Arithmetic Teacher 4:45-49, April, 1957.
- Rappaport, David. "An Investigation of the Degree of Understanding of Meanings in Arithmetic of Pupils in Selected Elementary Schools." unpublished doctoral dissertation, Northwestern University, Evanston, Illinois, 1957.
- Rappaport, David. "Testing for Meanings in Arithmetic," Arithmetic Teacher, 6:140-143, April, 1959.
- Rea, Robert E. and Reys, Robert E. "The Comprehensive Mathematics Inventory: An Experimental Instrument for Assessing Youngsters Entering School," Journal of Educational Measurement 7:45-47, Spring 1970.

- Romberg, Thomas A. "Contemporary Mathematics Test Series," Journal of Educational Measurement 5:349-351, Winter, 1966.
- Shulman, Lee S. "Psychology and Mathematics Education". The Sixty-Ninth Yearbook of the National Society for the Study of Education. NSSE, Chicago, 1970.
- Shuster, Albert H., and Pigge, Fred L., "Retention Efficiency of Meaningful Teaching," The Arithmetic Teacher 12:24 - 31, January, 1965.
- Sobel, Max A. and Johnson, Donovan A. "Analysis of Illustrative Test Items" Evaluation in Mathematics Twenty-Sixth Yearbook. National Council of Teachers of Mathematics. Washington, D.C. 1961.
- Spitzer, Herbert F. "Procedures and Techniques for Evaluating the Outcomes of Instruction in Arithmetic" Arithmetic 1948. G.T. Buswell ed. The University of Chicago Press, Chicago, 1948.
- Stokes, C. Newton, "80,000 Children's Reactions to Meanings in Arithmetic," The Arithmetic Teacher 5:281-286, December, 1958.
- Thiele, C.L. The Contribution of Generalization to the Learning of the Addition Facts. Bureau of Publications Teachers College, Columbia University, New York, 1938.
- Tyler, Ralph W., "Some Findings From Studies in the Field College Biology," Science 18: 133-42.
- Underwood, Benton J., "Laboratory Studies of Verbal Learning" in Theories of Learning and Instruction, Sixty-Third Yearbook of the National Society for the Study of Education, University of Chicago Press, Chicago, 1964.
- Van Engen, H., "Which Way Arithmetic," The Arithmetic Teacher 2:131-140, December, 1955.
- Weaver, J. Fred, "Evaluating and the Classroom Teacher." The Sixty-Ninth Yearbook of the National Society for the Study of Education. NSSE, Chicago, 1970.
- Weaver, J. Fred, "Some Areas of Misunderstanding About Meanings in Arithmetic." Elementary School Journal 51:35-41, September, 1950.
- Welch, Wayne W. "Curriculum Evaluation" Review of Educational Research 39:429-443, October, 1969.
- Wood, Dorothy (Adkins). Test Construction; Development and Interpretation of Achievement Tests. C.R. Merril Books. Columbus, Ohio 1960.

APPENDIX A

LIST OF JUDGES

LIST OF JUDGES

Mr. Richard Daly

Graduate student, elementary mathematics education.

Mr. Robert Dargavel

Graduate student, elementary mathematics education.

Mr. Frank Riggs

Graduate student, elementary mathematics education.

Dr. K. A. Neufeld,

Associate Professor, Department of Elementary Education,
University of Alberta.

APPENDIX B

OBJECTIVES OF THE MATHEMATICS PROGRAMS USED IN THIS STUDY

OBJECTIVES OF THE AW1 PROGRAM

Number Numeration (N)

1. Matches sets to determine more, less, and equivalence of sets and can find a set having one more element than a given set. (0-18)
2. Counts ordered and unordered sets and subsets of objects to determine the number property of the set. (0-18)
3. Given a written numeral, child circles the correct number of objects in a set to illustrate the given numeral. (0-18)
4. Children can write the standard numerals. (0-99)
5. Can tell which of two given numerals represents the larger number (or smaller). Also given two sets children determine the number property of each set and write a phrase to indicate the relationship. $a < b$; $a > b$.
6. Given a structured set of objects to 99 the student notates (a) tens and (b) ones and abstracts to the standard numeral form. (ab)
7. Given a number named as (a) tens and (b) ones the student writes the standard numeral and vice versa.
8. Student tells which number goes between two given numbers and which numbers come immediately before and after a given number.
9. Count by 1's to 99.
10. Skip count by 2, 3, 4, 5 to 99.

Addition and Subtraction (A)

1. Recognizes additive and subtractive action (union of sets,

removal of a subset). Can determine the number property of both sets and the union set or remainder set from appropriate arrays of numerals.

2. Given pictured unions and removals the student selects the correct number phrase to indicate that union or removal from an array of phrases.
3. Knows number facts sums and minuends to 18.
4. Given a number line and the solution to a problem the child completes a given number sentence to indicate the solution to the problem.
5. Can complete equations of the form $a + \square = C$; $a - \square = c$,
 $+ a = c$.
6. Can illustrate the relationship between addition and subtraction by completing equations of the form.
7. Given a partitioned set the child realizes four ways to illustrate that partitioning

$\square + \triangle = 0$	$0 - \triangle = \square$
$\triangle + \square = 0$	$0 - \square = \triangle$
8. Can complete number sentences of the form $a + b \ 0 \ C$; $a + b \ 0 \ C \ c + d$ by replacing 0 with one of $<$, $>$, $=$ to make a true sentence.
9. Adds three single digit number in two ways to illustrate the associative principle for addition. Can complete equations of the form $(2 + 3) + 1 = 2 + (- + -)$.

Money (Mo)

1. Child recognizes penny, nickel and dime and realizes the value of larger denomination coin in terms of those of smaller denomination.
2. Can find the value of a set of coins to 99¢.
3. Given the price of two objects child finds the total price (to 18¢)

Fractions (F)

1. Given a pictured set of objects the student partitions the set into halves or quarters.
2. Can choose a whole object divided into halves or quarters.

Time

1. Can tell time to the half hour.

Measurement (Me)

1. Children measure objects in inches and centimeters.
2. Liquid measure is introduced and the children realize the number of cups in two pints and a quart and the number of pints in a quart.
3. Can determine which of two pictured volumes given in pints, cups and quarts is larger.

Geometry

1. Child recognizes the triangle, square, circle, rectangle and hexagon.
2. Recognizes points on the inside and outside of a closed curve.

OBJECTIVES OF THE STAI PROGRAM

Number and Numeration (N)

1. Can determine, in a given situation, if a set of objects consists of many objects or a few objects. (many, few)
2. Can match equivalent sets and realizes they have the same number property.(0-9)
3. Knows word names and Hindu-Arabic numerals.(0-9)
4. Becomes familiar with the ideas of less than, greater than, equal to (does not use the symbols $<$, $>$, just $=$) i.e. matches sets to determine more, less and equivalence.(0-9)
5. Recognizes that a set has one more object than another.(is one greater)
6. Can order the counting numbers 1-10.
7. Can determine that number or numbers between two given numbers.
8. Realizes and can use numbers to indicate the position of an object in an arranged set of objects. (first-tenth)
9. Can use a pair of numbers to describe the location of an object in a chart.
10. Given sets of objects to 99 objects, the student realizes that he can group the objects by tens. He learns to record the number of tens and number of ones by using tally marks or numerals in tally plates.
11. Given a structured set of objects to 99, the student notates as "a tens and b ones" and abstracts to standard numeral form.

12. Given a standard numeral to 99 the student says the word name for that number.
13. Counts by 5's to 95 and by 10's to 90.
14. The ideas of betweenness of numbers, greater than, less than, and one greater than of the number section are extended to numbers to 99.

Addition and Subtraction (A)

1. Learns to recognize additive action and subtractive action.
2. Given pictured additive and subtractive actions the student writes the number phrase for each example (words and symbols).
3. Recognizes that number phrases illustrating the commutative property arise from different physical situations but represent the same number, i.e., realizes $a + b = b + a$.
4. Can complete equations of the form $\square + 3 = 3 + 2$ and $\triangle + 4 = \square + \triangle$ to illustrate the commutative property.
5. Solves addition problems sums to eight and subtraction questions with minuend to eight.
6. Child learns that a number can be named in many different ways (eg. $8 = 4 + 4$, $8 = 6 + 2$ etc. or $8 = 9 - 1$ but does not write $4 + 4 = 6 + 2$).
7. Learns that what can be done by a joining action can be undone by a separation action. Does not write eqns of the form $a + b = c$, $a = c - b$.

Money (Mo)

1. Can recognize penny, nickel, dime, quarter and has an awareness for the value of each coin.
2. Finds the value of sets of coins up to 99¢.
3. Can select from a collection of coins a combination of coins that is worth a specified amount (from sets of coins worth up to 99¢).
4. Can use pennies, nickels and dimes to make change for amounts less than 50¢.

Geometry (G)

1. Recognizes paths that represent closed and open curves.
2. Recognizes that a closed curve has an inside and outside (interior and exterior) and that an open curve has neither.
3. Recognizes one point between two points on a simple open curve (betweenness).

Measurement (Me)

1. Uses a non standard unit of length to measure objects.
2. Uses a non standard unit of capacity to measure volume.

OBJECTIVES OF THE AW2 PROGRAM

Number and Numeration (N)

1. Matches sets to determine more, less, equivalence of sets and also to find sets that have one more element than a given set.
2. Counts ordered and unordered sets and subsets of objects to determine the number property of the set.

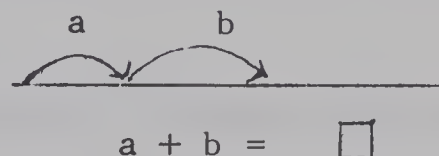
3. Given a written numeral, child circles the correct number of objects in a set to illustrate the given number.
4. Children can write standard numerals to 9,999.
5. Can tell which of two given numerals represents the larger number (0-9,999).
6. Given two sets, children determine the number property of each set and write a phrase to indicate the relationship between these numbers - $a < b$ or $a > b$.
7. Given a structured set of objects to 9,999 the student notates (a) thousands, (b) hundreds, (c) ones, and (d) units and abstracts to the standard numeral form - abcd.
8. Given a number named as (a) thousands, (b) hundreds, (c) tens, (d) ones, the student writes the standard numeral and vice versa.
9. Student tells which number or numbers goes between two given numbers, come immediately before and immediately after a given number (to 9,999).
10. Illustrates the ability to count to 10,000 by filling in blanks in short sequences of numbers to 9,999.
11. Skip counts by 2, 3, 4, 5, 10 to 9,999.
12. Can pick out odd and even numbers.
13. Can order a set of numbers from smallest to largest.

Addition and Subtraction (A)

1. Can do two-digit addition and subtraction problems with regrouping and do 4-digit addition and subtraction problems with no regrouping (sums to 9,999).
2. Can do 1 step addition and subtraction problems using all facts

of objective one. Child writes the number sentence and solves.

3. Understands the use of parenthesis in equations and is aware of the associative property of addition.
4. Can complete number sentences of the form $a + b \bigcirc c + d$ by inserting $<, >, =$ to make a true sentence (sums to 9,999 as indicated above).
5. Realizes the commutative property of addition and completes eqns. of the form $a + b = b + \square$.
6. Realizes the relationship between addition and subtraction and using the addition fact to find the subtraction answer $2 + 4 = \square$; $6 - 4 = \triangle$.
7. Given a number line and the solution to a problem child completes the corresponding number sentence.



8. Adds three single digit numbers in two different ways to illustrate the associative principle for addition. Can complete eqn. of the form. $(a + b) + c = a + (\quad + \quad)$.

Multiplication and Division (M)

1. Given 'a' sets of 'b' objects the children determine the total number of objects.
2. Given pictured equivalent sets and the eqn $a + b = \square$ children complete the equation ($a, b = 5$).
3. Children use the number line to illustrate multiplication as repeated addition. Given a solution showing 'a' added 'b' times on a number line child completes the following equation.

$$a + a + \dots + a = \square \qquad b \times a = \square$$

4. Does one step word problems. Child write number sentence $a \times b = \square$ and solves.
5. Given pictured sets divided into equivalent subsets the child completes equations of the form $a \div b = \square$.

Fractions (F)

1. Given a pictured set of objects the student partitions them into halves, thirds, quarters, fifths.
2. Given a pictured object child can tell whether $1/2$, $1/4$, $1/5$, $1/3$, $2/3$, $3/4$ of the object is a certain colour.

Measurement (Me)

1. Children measure objects in inches and centimeters, i.e., measure to the nearest inch and centimeter.
2. Children realize how many cups make a quart, how many pints make a quart.
3. Children can determine if a pictured volume in quarts, pints, cups is more, less, or equal to a given pictured volume.

Time (T)

1. Children can tell time to five minutes and notate it as 4.10, 4.25, 4.55, etc.

Money (Mo)

1. Children can recognize all Canadian coins up to the fifty cent piece and give their value in terms of smaller coins.
2. Child can make change for articles less than \$1.00.
3. Child can choose a set of coins of a certain value from a given

set of coins.

Geometry (G)

1. Child can distinguish open and closed curves and curves made up of straight sections.
2. Can distinguish between simple closed curves and those that are not simple.
3. Can identify and name line segments.
4. Recognize square, rectangle, circle, triangle and hexagon.
5. Is familiar with the notion of interior and exterior of simple closed curves and polygon.
6. Can match plane figures of the same size and shape.

OBJECTIVES OF THE STA2 PROGRAM

Number and Numeration (N)

1. Child can match equivalent sets and realizes they have the same number property (0-9) .
2. Becomes familiar with the ideas of betweenness, less than, equal to, greater than of the numbers (0-9)

(This is mainly review and these ideas are extended in the numeration section.)
3. Recognizes that a set has one more object than another (is one greater).
4. Can use numbers to indicate the position of an object in an arranged set of objects (first - 999th).
5. Can use a pair of numbers to describe the location of an object in a chart.

6. Given a structured set of objects to 999 the student notates as 'a' hundreds, 'b' tens, 'c' ones and abstracts to standard numeral form.
7. Given a standard numeral to 999 the student says the word name for that number.
8. Counts by 5's, 10's, 25's to any number < 1000 .
9. The ideas of betweenness of numbers, greater than, less than, and one greater than of the number section are extended to 999.
10. Given a standard numeral to 999 the student writes the numeral in expanded form, eg., $abc = (a) \text{ hundreds } (b) \text{ tens } (c) \text{ units}$.

Addition and Subtraction (A)

1. Learns to recognize additive action and subtractive action.
2. Given pictured additive and subtractive actions the student writes the number phrase for each example (words and symbols) (0-18).
3. Knows addition facts to 18 and subtraction facts with minuends to 18.
4. Recognizes that number phrases illustrating the commutative property arise from different physical situations but represent the same number, i.e. realizes $a + b = b + a$.
5. Can complete equations of the form $\square + 3 = 3 + 2$ and $\square + 4 = \square + \triangle$ to illustrate the commutative property.
6. Child learns that a number can be named in many different ways (e.g. $8 = 4 + 4$, $8 = 6 + 2$ etc. or $8 = 9 - 1$ and now writes egns at the form $4 + 4 = 6 + 2$
 $1 + 2$ is less than $2 + 2$
 $4 + 4$ is greater than $4 + 2$
 note: no symbols for greater than or less than.

7. Child extends his knowledge of the relationship between addition and subtraction. Given pictured additive and subtractive actions to illustrate $a + b = c$ and $c - b = a$ writes the appropriate number phrase for each picture.
8. Given one-step addition and subtraction problems in words illustrated by pictures the child (a) writes the appropriate number phrase (b) writes the appropriate equation needed to find the answer using a placeholder for the unknown (c) solves the problem eg. eqn's of the form

$$a + b = \square$$

$$a - b = \square$$

$$\square - b = a$$

$$\square + b = a$$
9. Can do addition questions with three addends (sums to 18).
10. Recognizes the associative property of addition. Given a phrase of the form $a + b + c$ the child writes an equation to illustrate the associative property of addition and works out both sides of the equation.

Multiplication and Division (M)

1. Recognizes the joining action (bringing together of equal sets) that suggests multiplication and the separating action (separating into equal sets) that suggests division.
2. Given illustrated multiplication or division actions describes the action in any of the following ways:

a times b
 a b's
 b multiplied by a
 b divided by a
 $a \times b$
 $b \div a$

3. Knows the basic multiplication facts to 18 and the division facts with dividends to 18.
4. Recognizes more ways of naming the same number eg. $12 - 2$, 2×3 , $7 - 1$, $3 + 3$.
5. Child learns the relationship between multiplication and division, i.e., given a pictured multiplication action the child can pick out the correct corresponding division action.
6. Given pictured equal sets containing a subset with a particular attribute the child makes a statement expressing the ratio of the number objects with this attribute to the total number in the set eg. 1 out of 3 were blue.
7. Given pictures of different sets writes statements of the ratio of the number in one set to the number in another set. (recognizes instances when such a comparison may not be made.

Money (Mo)

1. Recognizes a fifty cent piece and a one dollar bill and knows their value in terms of coins of lesser amounts. eg. $50¢ = 5 \text{ dimes or } 10 \text{ nickels}$.
2. Finds the value of sets of coins up to $99¢$.
3. Can select from a collection of coins a combination of coins that is worth a specified amount. (from sets of coins worth up to $100¢$)
4. Makes change for amounts that are less than $\$1.00$.

Geometry (G)

1. Recognizes paths that represent closed and open curves.

2. Recognizes that a closed curve has an inside and outside (interior and exterior) and that an open curve has neither.
3. Can distinguish between a line and a curve.
4. Recognize one point between two points on a simple open curve or a line.(betweeness)
5. Can identify and name line segments.
6. Can identify lines that intersect and names points of intersection.
7. Realizes that two points determine a line.
8. Learns that a polygon is a closed curve made up entirely of line segments. Can name the sides of the polygon.

Measurement (Me)

1. Uses a ruler to measure lengths in standard units of inches and feet.
2. Is introduced to pint and quart as standard units for measure of capacity.
3. Learns that 2 pints make a quart.
4. Given pictures showing various volumes is able to tell whether a pictured volume (both given in terms of pints and quarts) is equal to less than, greater than the given volume.

APPENDIX C

JUDGES MATCHINGS OF ITEMS TO OBJECTIVES IN THE ASSESSMENT
OF THE CONTENT VALIDITY OF THE TESTS USED IN THIS STUDY

CONTENT VALIDITY RESULTS COOPERATIVE 12A TEST VS AW1 PROGRAM

Item	Judges				Valid		
	1	2	3	4		A	D
1	N2	N2	N2	N1	Y	3	3
2	N1	N1	N1	N1	Y	6	0
3		N1	N2	N1	Y	1	5
4	N7	N7	N7	N7	Y	6	0
5			N2	N2	Y	2	4
6		N1	N2		Y	1	5
7		N2	N2		Y	2	4
8			N2		N	3	3
9			N2	N2	Y	2	4
10	N8	N8	N8	N8	Y	6	0
11					N	6	0
12		F1		F2	Y	1	5
13		F1	F1	F2	Y	1	5
14	N5	A8	A3	N1	Y	0	6
15					N	6	0
16					N	6	0
17	A3	A3	A3	A1	Y	3	3
18	A3		A3	A9	Y	1	5
19				N1	N	3	3
20				A1	N	3	3
21					N	6	0
22	N5	A8	A8	N5	Y	1	5
23					N	6	0
24	A2	A1	A1	A1	Y	3	3
25	A2	A1	A1	A1	Y	3	3
26	A2	A1	A1	A1	Y	3	3
27					N	6	0
28					N	6	0
29		Mo2	Mo2		Y	2	4
30	Mo1	Mo1	Mo1	Mo1	Y	6	0
31					N	6	0
32	T1	T1	T1	T1	Y	6	0
33	Me1	Me1	Me1	Me1	Y	6	0
34					N	6	0
35					N	6	0
36					N	6	0
37					N	6	0
38		G1	G1		Y	2	4
39		G1	G1	G1	Y	3	3
40					N	6	0
41					N	6	0
42	N1	N1	N1	N1	Y	6	0
43	N2	N3	N6	N1	Y	0	6
44	N7	N7	N6	N6	Y	2	4
45	N1	N1	N1	N1	Y	6	0

CONTENT VALIDITY RESULTS COOPERATIVE 12A TEST VS AW1 PROGRAM (contd.)

Item	Judges				Valid	
	1	2	3	4	A	D
46	N5	N4	N5	AS3	Y	1
47					N	6
48	N1		N1	N1	Y	3
49	A4	A4	A4	A4	Y	6
50			N1	N1	Y	2
51		A7	A3	A5	Y	0
52	A3	A5	A5	A5	Y	3
53	Mo1			Mo2	Y	1
54		G1	G1		Y	2
55					M	6

Key	N - Number Numeration	G - Geometry
	A - Addition and Subtraction	Me - Measurement
	M - Multiplication	Mo - Money
	F - Fractions	T - Time

A - Agreements

D - Disagreements

Valid Items 36

Total A 207

Total D 123

Judge Reliability .77

CONTENT VALIDITY RESULTS COOPERATIVE 12A TEST VS STAI PROGRAM

Item	Judges				Valid		
	1	2	3	4		A	D
1	N3	N3	N4	N3	Y	3	3
2	N4	N4	N1	N1	Y	2	4
3	N2	N4	N2	N3	Y	1	5
4	N11	N11	N12	N12	Y	2	4
5	N8	N8	N8	N8	Y	6	0
6					N	6	0
7	N8	N2		N3	Y	0	6
8	N8		N8	N8	Y	3	3
9	N8		N8	N8	Y	3	3
10	N7	N7	N7	N7	Y	6	0
11					N	6	0
12					N	6	0
13					N	6	0
14	N6	N4	N13	AS5	Y	0	6
15					N	6	0
16					N	6	0
17	A1	A5	A1	A5	Y	2	4
18	A2		A1	A2	Y	1	5
19					N	6	0
20			A1		N	3	3
21					N	6	0
22	A2			A5	Y	1	5
23	A6				N	6	0
24	A2	A2	A5	A5	Y	2	4
25	A2	A2	A5	A5	Y	2	4
26	A2	A2	A5	A5	Y	2	4
27					N	6	0
28					N	6	0
29	Mo1	Mo2	N14	Mo2	Y	1	5
30		Mo1	Mo1	Mo1	Y	3	3
31					N	6	0
32					N	6	0
33	Me1	Me1	Me1	Me1	Y	6	0
34					N	6	0
35	Me2				N	3	3
36					N	6	0
37			Me1		N	3	3
38	G1				N	3	3
39					N	6	0
40			Me1		N	3	3
41					N	6	0
42	N2	N2	N2	N2	Y	6	0

CONTENT VALIDITY RESULTS COOPERATIVE 12A TEST VS STA1 PROGRAM (contd.)

Item	Judges				Valid		
	1	2	3	4		A	D
43	N11	N10	N10	N11	Y	2	4
44	N11	N11	N11	N11	Y	6	0
45	N2	N2	N2	N2	Y	6	0
46	N3	N11	N4	N4	Y	1	5
47					N	6	0
48			N2	N2	Y	2	4
49	A2		A2	A5	Y	1	5
50			N2	N2	Y	2	4
51	A4	A4	A3	A5	Y	1	5
52			A5		N	3	3
53	Mo3		Mo3		Y	2	4
54					N	6	0
55					N	6	0

Key N - Number Numeration G - Geometry
 A - Addition and Subtraction Me - Measurement
 M - Multiplication Mo - Money
 F - Fractions T - Time

A - Agreements

D - Disagreements

Valid Items 29

Total A 213

Total D 117

Judge Reliability .79

CONTENT VALIDITY RESULTS COOPERATIVE 12A TEST VS AW2 PROGRAM

Item	Judges				Valid		
	1	2	3	4		A	D
1	N2	N2	N2	N1	Y	3	3
2	N1	N1	N1	N1	Y	6	0
3		N2	N2	N2	Y	1	5
4	N8	N8	N8	N8	Y	6	0
5			N2	N2	Y	2	4
6		N1	N2	N13	Y	0	6
7		N2	N2		Y	2	4
8			N2		N	3	3
9			N2	N2	Y	2	4
10	N9	N9	N9	N9	Y	6	0
11	N12	N12	N12	N12	Y	6	0
12	F1	F1		F2	Y	1	5
13	F1	F2	F1	F2	Y	2	4
14	AS4	A4	AS2	N1	Y	1	5
15	N4	N4	N8	N8	Y	2	4
16		F2		F2	Y	2	4
17	AS1	A1	A1	A1	Y	6	0
18		A8	A2	A8	Y	2	4
19	M1	M1	M1	M1	Y	3	3
20	A2	A1	A1	A1	Y	2	4
21	M4	M4	M2	M2	Y	2	4
22	A4	A4	A4	A1	Y	3	3
23	M4	M2	M1	M1	Y	1	5
24	A2	A2	A2	A2	Y	6	0
25	A2	A2	A2	A2	Y	6	0
26	A2	A2	A2	A2	Y	6	0
27		M2	M1	M4	Y	0	6
28				M11	N	3	3
29		Mo1	Mo2	N5	Y	0	6
30	Mo1	Mo1	Mo1	Mo1	Y	6	0
31					N	6	0
32	T1	T1	T1	T1	Y	6	0
33	Me1	Me1	Me1	Me1	Y	6	0
34					M	6	0
35				Me2	N	3	3
36					M	6	0
37					N	6	0
38		G4	G4		Y	2	4
39		G4	G4	G4	Y	3	3
40					N	6	0
41					N	6	0
42	N1	N1	N1	N1	Y	6	0
43	N2	N3	N7	N1	Y	0	6

CONTENT VALIDITY RESULTS COOPERATIVE 12A TEST VS AW2 PROGRAM (contd.)

Item	Judges				Valid	
	1	2	3	4	A	D
44	N7	N8	N7	N7	Y	3
45	N1	N1	N1	N1	Y	6
46		N4	N6	N4	Y	1
47	A1		A2		Y	1
48	N1	N1	N1	N1	Y	6
49	A7	A7	A7	A7	Y	6
50		N1	N1	N1	Y	3
51	A5	A5	A2	A5	Y	3
52	A1	A2	A2	A6	Y	1
53	Mo1	Mo1	Mo1	Mo2	Y	3
54		G4	G4	G4	Y	3
55				G4	N	3

Key N - Number Numeration G - Geometry
 A - Addition and Subtraction Me - Measurement
 M - Multiplication Mo - Money
 F - Fractions T - Time

A - Agreements

D - Disagreements

Valid Items 45

Total A 192

Total D 138

Judge Reliability .74

CONTENT VALIDITY RESULTS COOPERATIVE 12A TEST VS STA2 PROGRAM

Item	Judges				Valid		
	1	2	3	4		A	D
1	N1	N1	N1	N1	Y	6	0
2	N1	N2	N1	N2	Y	2	4
3	N1	N1	N1	N2	Y	3	3
4	N10	N6	N10	N10	Y	3	3
5	N4	N4	N4	N4	Y	6	0
6			N4		N	3	3
7	N4	N2	N4		Y	1	5
8	N4		N4	N4	Y	3	3
9	N4		N4	N4	Y	3	3
10	N2	N2	N2	N9	Y	3	3
11					N	6	0
12					N	6	0
13					N	6	0
14	N2	N2	AS3	N9	Y	1	5
15	N7	N7	N7	N7	Y	6	0
16					N	6	0
17	A3	A2	A3	AS1	Y	1	5
18	A1	A10	A9	AS1	Y	1	5
19			M2	N1	Y	1	5
20				AS1	N	3	3
21	M2		M2	AS1	Y	1	5
22	A3	A6	A3	A1	Y	1	5
23	M3		M3	M2	Y	1	5
24	A2	A2	A2	A8	Y	3	3
25	A2	A2	A2	A8	Y	3	3
26	A2	A2	A2	A8	Y	3	3
27			M2	M6	Y	2	4
28	M6				N	3	3
29	Mo2	Mo2	Mo2	Mo2	Y	6	0
30	Mo1	Mo1	Mo1	Mo1	Y	6	0
31					N	6	0
32					N	6	0
33	Me1	Me1	Me1	Me1	Y	6	0
34					M	6	0
35					N	6	0
36					M	6	0
37				Me1	M	3	3
38				G5	M	3	3
39					M	6	0
40				Me1	M	3	3
41				AS6	M	3	3
42	N1	N1	N1	N1	Y	6	0
43	N6	N6	N6	N6	Y	6	0

CONTENT VALIDITY RESULTS COOPERATIVE 12A TEST VS STA2 PROGRAM (contd.)

Item	Judges				Valid		
	1	2	3	4		A	D
44	N6	N6	N6	N10	Y	3	3
45	N1	N1	N1	N1	Y	6	0
46	N1	N2	N2	A5	Y	1	5
47				A6	M	3	3
48		A3	N1	N1	Y	1	5
49	A6		A3	A6	Y	1	5
50			N1	N1	Y	2	4
51	A4	A5	A3	A4	Y	2	4
52	A3	A3	A3	A8	Y	3	3
53	Mo1	Mo1	Mo1	Mo1	Y	6	0
54					N	6	0
55					N	6	0

Key N - Number Numeration G - Geometry
 A - Addition and Subtraction Me - Measurement
 M - Multiplication Mo - Money
 F - Fractions T - Time

A - Agreements

D - Disagreements

Valid Items 35

Total A 205

Total D 125

Judge Reliability .77

CONTENT VALIDITY RESULTS SRA TEST VS AW1 PROGRAM

Item	Judges				Valid		
	1	2	3	4		A	D
A		G1			N	3	3
B					M	6	0
C	F2	F2	F2	F2	Y	6	0
D			M2		N	3	3
E	N2				M	3	3
F	A2	A2	A2	A1	Y	3	3
G	A3	A3	A2	A1	Y	1	5
H					N	6	0
I	N5	N4	N4	N5	Y	2	4
J	A3	A3	A3	M7	Y	3	3
K					N	6	0
L	A6	A6	A7	A7	Y	2	4
M	Mo2	Mo2	Mo2	Mo2	Y	6	0
N	N6	N6	N6	N6	Y	6	0
O					N	6	0
P					N	6	0
Q					N	6	0
R	N10	A10	N10		Y	1	5
S			Me1	Me2	Y	1	5
T	A8	A8	A8	A8	Y	6	0
U					N	6	0
V					N	6	0
W	A9	A9	A9	A9	Y	6	0
X			N5		N	3	3
Y	N7			N7	Y	2	4
Z	G1	G1	G1	G1	Y	6	0
AA					N	6	0
BB					N	6	0
CC					N	6	0
DD					N	6	0
EE				As8	N	3	3
FF			A3	A8	Y	1	5
GG	A5	A9	A5	A6	Y	1	5
HH					N	6	0
II					N	6	0
JJ					N	6	0
KK	A9		A9	A9	Y	3	3
LL				N7	N	3	3
MM				A6	M	3	3
NN			A3	A9	Y	1	5

CONTENT VALIDITY RESULTS SRA TEST VS AW1 PROGRAM (contd.)

Item	Judges				Valid	
	1	2	3	4	A	D
OO					N 6	0
PP					N 6	0

Key N - Number Numeration G - Geometry
 A - Addition and Subtraction Me - Measurement
 M - Multiplication Mo - Money
 F - Fractions T - Time

A - Agreements

D - Disagreements

Valid Items 24

Total A 180

Total D 72

Judge Reliability .81

CONTENT VALIDITY RESULTS SRA TEST VS STA1 PROGRAM

Item	Judges				Valid	
	1	2	3	4	A	D
A					N	0
B					N	0
C					N	0
D	N8	N8	N8	N8	Y	0
E					N	0
F				AS1	N	3
G					N	0
H					N	0
I	N14	N14	N14		Y	3
J				AS6	N	3
K	N9	N9	N9	N9	Y	0
L		A6	A7		Y	5
M	Mo2	Mo2	Mo2	Mo2	Y	0
N	N11	N11	N10	N11	Y	3
O					N	0
P					N	0
Q					N	0
R					N	0
S					N	0
T					N	0
U					N	0
V					N	0
W					N	0
X	N14		N4		Y	5
Y				N11	N	3
Z					N	0
AA					N	0
BB	AS5				N	3
CC					N	0
DD					N	0
EE					N	0
FF					N	0
GG					N	0
HH					N	0
II					N	0
JJ					N	0
KK					N	0
LL					N	0
MM					N	0
NN					N	0
OO					N	0
PP					N	0

CONTENT VALIDITY RESULTS SRA TEST VS STA1 PROGRAM (contd.)

Key	N - Number Numeration	G - Geometry
	A - Addition and Subtraction	Me - Measurement
	M - Multiplication	Mo - Money
	F - Fractions	T - Time

A - Agreements

D - Disagreements

Valid Items 7

Total A	224
Total D	28
Judge Reliability	.94

CONTENT VALIDITY RESULTS SRA TEST VS AW2 PROGRAM

Item	Judges				Valid		
	1	2	3	4		A	D
A		G4			N	3	3
B	N5	N5	N5	N5	Y	6	0
C	F2	F2		F2	Y	3	3
D			N2	N2	Y	2	4
E	G3	G3		G3	Y	3	3
F		A1	A2		Y	1	5
G	A1	A1	A2		Y	1	5
H	M1	M4	M1		Y	1	5
I	N5	N5	N5	N5	Y	6	0
J	A1	A1	A2	N8	Y	1	5
K					N	6	0
L	A6	A6	A6	A6	Y	6	0
M		Mo3		Mo3	Y	2	4
N	N7	N7	N7	N8	Y	3	3
O	N8	N8	N8	N8	Y	6	0
P	T1	T1		T1	Y	3	3
Q					N	6	0
R	N11	N11	N11	N11	Y	6	0
S				Me1	N	3	3
T	A4	A4	A4	A4	Y	6	0
U				T1	N	3	3
V					N	6	0
W	A8	A8	A8	A3	Y	3	3
X			N13	N8	Y	1	5
Y	N8	AS1	AS1	N8	Y	2	4
Z	G4	G4		G4	Y	3	3
AA					N	6	0
BB	M4	M4	M1	M1	Y	2	4
CC	M5		M1	D1	Y	1	5
DD	Mo2	Mo2	A2		Y	1	5
EE		A1	A2	A1	Y	1	5
FF	A6	A1		A1	Y	1	5
GG	A6	A1	A2	A1	Y	1	5
HH	A6	A1	A1	A1	Y	3	3
II		A1	A2	A1	Y	1	5
JJ				M4	N	3	3
KK	A3	A8	A3	A3	Y	3	3
LL		A1	A1	N8	Y	1	5
MM	A6	A1	A1	A1	Y	3	3
NN	A3		A3	A3	Y	3	3
OO	A1	A1	A1	A1	Y	6	0
PP				M4	N	3	3

CONTENT VALIDITY RESULTS SRA TEST VS AW2 PROGRAM (contd.)

Key	N - Number Numeration	G - Geometry
	A - Addition and Subtraction	Me - Measurement
	M - Multiplication	Mo - Money
	F - Fractions	T - Time

A - Agreements

D - Disagreements

Valid Items 33

Total A 131

Total D 121

Judge Reliability .68

CONTENT VALIDITY RESULTS SRA TEST VS STA2 PROGRAM

Item	Judges				Valid	
	1	2	3	4	A	D
A					N	6
B	N9	N9	N9	N9	Y	6
C					N	6
D	N4	N4	N4	N4	Y	6
E		G5	G5	G5	Y	3
F	A2	A3	A2	A1	Y	1
G	A3	A3	A2	A6	Y	6
H	M2	M2	M2	M2	Y	6
I	N9	N9	N9	N9	Y	6
J	A6	A6	A3	N9	Y	1
K	N5	N5	N5	N5	Y	6
L	A8	A7	A7	A7	Y	3
M	Mo2	Mo2	Mo2	Mo2	Y	6
N	N1	N1	N1	N10	Y	3
O	N10	N10	N10	N10	N	6
P					N	6
Q					N	6
R					N	6
S	Me1	Me1		Me1	Y	3
T	N9		A6	A3	Y	0
U					N	6
V					N	6
W	A10	A10	A10		Y	3
X	N9		N2		Y	1
Y				N7	N	3
Z					N	6
AA					N	6
BB	M2	M3	M2	M4	Y	1
CC	M3	M3	M2		Y	1
DD					N	6
EE					N	6
FF	A3	A3	A3		Y	3
GG	A3	A3	A3	A3	Y	6
HH					N	6
II					N	6
JJ					N	6
KK	A9	A9	A9	A9	Y	6
LL					N	6
MM					N	6
NN			A3		N	3
OO					N	6
PP	M3		M3		Y	2

CONTENT VALIDITY RESULTS SRA TEST VS STA2 PROGRAM (contd.)

Key	N - Number Numeration	G - Geometry
	A - Addition and Subtraction	Me - Measurement
	M - Multiplication	Mo - Money
	F - Fractions	T - Time

A - Agreements

D - Disagreements

Valid Items 23

Total A	188
Total D	64
Judge Reliability	.85

Table 19

RESULTS OF THE JUDGES' ASSESSMENT OF THE CONTENT VALIDITY OF THE TESTS USED IN THIS STUDY

Test and Program	Content Validity	Total Agreements	Total Disagreements	Reliability $R = \frac{2a}{2a+d}$	Valid Items
Cooperative Test on the AW1 Program	16/30 (.53)	207	123	.77	36
SRA Test on the AW1 Program	15/30 (.50)	180	72	.81	18
Cooperative Test on the STAl Program	17/30 (.57)	215	115	.79	29
SRA Test on the STAl Program	6/30 (.20)	224	28	.94	7
Cooperative Test on the AW2 Program	23/41 (.56)	192	138	.74	45
SRA Test on the AW2 Program	18/41 (.44)	131	121	.68	33
Cooperative Test on the STA2 Program	18/43 (.42)	205	125	.77	35
SRA Test on the STA2 Program	17/43 (.40)	188	64	.85	23

APPENDIX D

TABLES SHOWING UNSUITABLE ITEMS AND THEIR DIFFICULTY INDEXES

Table 20A
UNSUITABLE ITEMS AND THEIR DIFFICULTY INDEXES COOPERATIVE TEST

Gr. 1 Sample Item	AW1		STAL		Gr. 2 Sample		AW2		STA2	
	Dif.	Item	Dif.	Item	Dif.	Item	Dif.	Item	Dif.	Item
11	.367	5	.733	1	.967	1	1.000	1	1.000	1
16	.583	9	.633	4	.900	4	1.000	4	1.000	4
23	.417	10	.900	6	.400	24	.850	8	1.000	10
25	.900	11	.300	11	.433	28	.967	20	.900	25
28	.950	19	.667	16	.500	31	1.000	22	.967	31
36	.183	23	.333	25	.867	32	.917	24	.833	32
44	.383	28	.933	28	.967	36	.250	25	.967	34
45	.967	31	.933	36	.200	53	.917	27	.967	36
		32	.933	38	.233			28	1.000	53
		35	.633	39	.833			29	.967	
		36	.167	42	.267			31	1.000	
		44	.333	45	.967			32	.933	
		45	.967	46	.600			33	.633	
		52	.759	47	.667			45	1.000	
				49	.700			47	.900	
				51	1.000			49	.833	
								53	.967	

Mean Difficulty Indexes of Unsuitable Items

Grade One Sample	.594	Grade Two Sample	.863
AW1	.659	AW2	.933
STAL	.656	STA2	.874

Table 20B
UNSUITABLE ITEMS AND THEIR DIFFICULTY INDEXES SRA TEST

Gr. 1		AW1		STAL		Gr. 2		AW2		STA2	
Sample	Item	Dif.	Item	Dif.	Item	Sample	Item	Dif.	Item	Dif.	Item
12	8	.200	1	.467	4	4	8	.717	4	1.000	4
14	14	.100	6	.533	17	17	17	.167	6	.200	6
17	16	.217	7	.467			23		17	.600	17
23	17	.400	9	.267			30			.800	
26	22	.517	12	.167							
31	23	.450	17	.233							
34	32	.153	18	.233							
36	34	.154	19	.300							
40	40	.216	20	.233							
41	41	.250	21	.100							
42	42	.250	23	.300							
			24	.233							
			25	.233							
			26	.700							
			31	.333							
			36	.115							
			40	.250							
			41	.316							
			42	.294							

Mean Difficulty Indexes of Unsuitable Items

Grade One Sample	.264	Grade Two Sample	.442
AW1	.326	AW2	.650
STAL	.278	STA2	.589

APPENDIX E

TABLES SHOWING THE ITEMS AND THEIR DIFFICULTY INDEXES ON THE
BASIC FACTS AND NUMBER NUMERATION SUBTESTS OF THE COOPERATIVE
AND SRA TESTS

Table 21A

A LISTING OF THE ITEMS FROM THE BASIC FACTS AND NUMBER-NUMERATION SUBTESTS OF THE COOPERATIVE TEST TOGETHER WITH THEIR DIFFICULTY INDEXES (GR. I).

Item	Basic Facts		Difficulty Index		Gr. I	Item	Number-Numeration		Gr. I
	AWI		STAI				AWI	STAI	
17	.267*		.333*		.300	1	.900*	.967*	.433
18	.700*		.433*		.567	2	.800*	.800*	.800
19	.667		.500		.583	3	.800*	.900*	.850
20	.433		.633		.533	4	.800*	.900*	.850
21	.633		.367		.500	5	.733*	.900*	.817
22	.600*		.567*		.583	6	.600*	.400	.500
23	.333		.500		.417	7	.567*	.500*	.533
24	.733*		.900*		.817	8	.767	.933*	.850
25	.933*		.867*		.900	9	.633*	.233*	.433
26	.833*		.467*		.650	10	.900*	.800*	.850
27	.933		.900		.917	11	.300*	.433	.367
47	.733		.667		.700	14	.700*	.833*	.767
48	.767*		.667*		.717	15	.467	.633	.550
49	.600*		.700*		.650	28	.933	.967	.950
50	.500*		.667*		.583	42	.433*	.267*	.350
51	.931*		1.000*		.966	43	.833*	.667*	.750
52	.759*		.467		.610	44	.333*	.433*	.383
						45	.967*	.967*	.967
						46	.433*	.600*	.517

* Indicates a valid item

Table 21A (continued)

Mean difficulty index for the basic facts subtest		
(a)	AW1	.668
(b)	STAL	.626
(c)	Gr.I	.647
Mean difficulty index for the Number-Numeration subtest		
(a)	AW1	.679
(b)	STAL	.691
(c)	Gr.I	.685
Content Validity		
Basic Facts		
	AW1 4/9	STAL 4/7
Number Numeration		
	AW1 6/10	STAL 8/14

Table 21B

A LISTING OF THE ITEMS FROM THE BASIC FACTS AND NUMBER-NUMERATION SUBTESTS OF THE COOPERATIVE TEST TOGETHER WITH THEIR DIFFICULTY INDEXES (GR.2).

Item	Basic Facts		Gr.2	Item	Number-Numeration		Gr.2
	Difficulty AW2	Index STA2			Difficulty AW2	Index STA2	
17	.400*	.500*	.450	1*	1.000*	1.000*	1.000
18	.833*	.767*	.800	2*	.867*	.900*	.883
19	.900*	.800*	.850	3*	.900*	.867*	.883
20	.900*	.833*	.867	4*	1.000*	1.000*	1.000
21	.633*	.500*	.567	5*	.900*	.867*	.883
22	.967*	.800*	.833	6*	.867	.533	.700
23	.633*	.600*	.617	7*	.900*	.667*	.783
24	.833*	.867*	.850	8	1.000	.933*	.967
25	.967*	.967*	.967	9	.933*	.667*	.800
26	.900*	.867*	.883	10	.900*	1.000*	.950
27	.967*	.967*	.967	11	.767*	.667	.717
47	.900*	.867	.883	14	.933*	.800*	.867
48	.833*	.767	.800	15	.700*	.667*	.683
49	.833*	.700*	.767	28	1.000	.933	.967
50	.833*	.767*	.800	42	.733*	.767*	.750
51	.967*	.933*	.950	43	.933*	.867*	.900
52	.800*	.867*	.833	44	.633*	.500*	.567
				45	1.000*	.900*	.950
				46	.833*	.767*	.800

* Indicates a valid item

Table 21B (continued)

Mean difficulty index for the Basic Facts Subtest

(a)	AW2	.829
(b)	STA2	.786
(c)	Gr. 2	.808

Mean difficulty index for the Number-Numeration Subtest

(a)	AW2	.884
(b)	STA2	.805
(c)	Gr.2	.845

Content Validity

Basic Facts

AW2	8/13
STA2	9/17

Number-Numeration

AW2	5/13
STA2	8/10

Table 21c

A LISTING OF THE ITEMS FROM THE BASIC FACTS AND NUMBER-NUMERATION SUBTESTS OF THE SRA TEST TOGETHER WITH THEIR DIFFICULTY INDEXES (GR. I).

Item	Basic Facts		Gr. I	Number-Numeration			Gr. I
	AW1	Difficulty Index STAI		Item	AW1	Difficulty Index STAI	
F (6)	.567*	.533	.550	B (2)	.700	.633	.667
G (7)	.267*	.167	.217	D (4)	.500	.533*	.517
H (8)	.933	.867	.900	I (9)	.433*	.267*	.350
J (10)	.467*	.200	.333	N (14)	.100*	.100*	.100
L (12)	.233*	.167*	.200	O (15)	.200	.400	.300
W (23)	.500*	.300	.400	R (18)	.400*	.233	.317
Y (25)	.100*	.233	.167	T (20)	.267*	.233	.250
AA (27)	.267	.200	.233	X (24)	.533	.233*	.383
BB (28)	.800	.700	.750				
CC (29)	.400	.500	.450				
EE (31)	.567	.333	.450				
FF (32)	.700*	.467	.583				
GG (33)	.633*	.400	.517				
HH (34)	0.000	.310	.153				
II (35)	.690	.577	.636				
JJ (36)	.192	.115	.154				
KK (37)	.440*	.385	.412				
LL (38)	.714	.522	.614				
MM (39)	.263	.318	.293				
NN (40)	.176*	.250	.216				
OO (41)	.176	.316	.250				
PP (42)	.200	.294	.250				

Table 21C (continued)

* Indicates a valid item

Mean difficulty index for Basic Fact Subtest

- (a) AW1 .422
- (b) STA1 .371
- (c) Gr.I .397

Mean difficulty index for the Number-Numeration Subtest

- (a) AW1 .392
- (b) STA1 .300
- (c) Gr.I .346

Content Validity

Fasic Facts

- AW1 9/9
- STA1 1/7

Number-Numeration

- AW1 6/10
- STA1 4/14

Table 21D

A LISTING OF THE ITEMS FROM THE BASIC FACTS AND NUMBER-NUMERATION SUBTESTS OF THE SRA TEST TOGETHER WITH THEIR DIFFICULTY INDEXES (GR. 2).

Basic Facts			Difficulty Index			Number-Numeration			
Item	AW2	STA2	Gr.2	Item	AW2	STA2	Gr.2		
F (6)	.800*	.800*	.800	B (2)	.867*	.667*	.767		
G (7)	.733*	.333*	.533	D (4)	.600*	.833*	.717		
H (8)	1.000*	.800*	.900	I (9)	.433*	.633*	.633		
J (10)	.867*	.733*	.800	N (14)	.567*	.500*	.533		
L (12)	.367*	.367*	.367	O (15)	.767*	.633	.700		
W (23)	.600*	.433*	.517	R (18)	.667*	.400	.533		
Y (25)	.667*	.367	.517	T (20)	.600*	.100*	.350		
AA (27)	.600	.500	.550	X (24)	.500*	.367*	.433		
BB (28)	.800*	.733	.767						
CC (29)	.767*	.533*	.650						
EE (31)	.700*	.724	.712						
FF (32)	.733*	.857*	.793						
GG (33)	.867*	.821*	.845						
HH (34)	.567*	.500	.534						
II (35)	.900*	.630	.772						
JJ (36)	.733	.630	.684						
KK (37)	.900*	.704*	.807						
LL (38)	.567*	.500	.536						
MM (39)	.767*	.600	.691						
NN (40)	.433*	.174	.321						
OO (41)	.733*	.409	.596						
PP (42)	.556	.571*	.563						

Table 21D (continued)

* Indicates a valid item

Mean difficulty index for the Basic Facts Subtest

(a)	AW2	.712
(b)	STA2	.578
(c)	Gr.2	.645

Mean difficulty index for the Number-Numeration Subtest

(a)	AW2	.650
(b)	STA2	.517
(c)	Gr.2	.584

Content Validity

Basic Facts

AW2	9/13
STA2	9/7

Number Numeration

AW2	7/13
STA2	4/10

APPENDIX F

THE COOPERATIVE PRIMARY TEST, MATHEMATICS, FORM 23A

CONTENT VALIDITY

Cooperative 23A

-

AW2 Program

Cooperative 23A						AW2 Program					
Objective	1	2	3	4	Tested	Objective	1	2	3	4	Tested
N 1	X	X			Y	M 1	X		X		Y
2		X	X		Y	2					N
3					N	3	X	X	X	X	Y
4	X	X			Y	4	X	X	X	X	Y
5	X	X	X	X	Y	5		X			N
6		X	X	X	Y	F 1		X	X		Y
7		X	X		Y	2	X	X	X	X	Y
8	X		X	X	Y	Me 1		X	X	X	Y
9					N	2	X		X	X	Y
10					N	3					N
11		X	X	X	Y	T 1		X	X		Y
12	X	X	X	X	Y	Mo 1					N
13					N	2	X	X	X	X	Y
A 1	X	X			Y	3					N
2			X	X	Y	G 1			X		N
3					N	2					N
4	X			X	Y	3					N
5					N	4	X	X	X	X	Y
6					N	5					N
7					N	6	X				N
8		X			N						

KEY

N - Number Numeration
 A - Addition and Subtraction
 M - Multiplication
 F - Fractions

 G - Geometry
 Me- Measurement
 Mo- Money
 T - Time

Valid Items 32
 Content Validity .52
 Judge Reliability .71

The Cooperative Primary Test, Mathematics, Form 23A

This is a sixty item test designed to be used as a post-test in grade two and as a pre-test in grade three. The handbook lists nine subtests, number, symbolism, operation, function and relation, approximation proof, measurement, estimation, geometry. The number of items on each of these subtests range from a high of sixteen on the operation subtest to a low of one on the estimation subtest.

The reliability of this test as reported in the handbook is .81 for grade two pupils writing in the spring and .81 for grade three pupils in the fall.

The correlation between the Cooperative Primary Test, Mathematics, Form 12A and this test is .71.

The content validity coefficient of this test with respect to the AW2 program is .52 with thirty-two valid items. With respect to the STA2 program the content validity coefficient is .44 with thirty-one valid items. The judge reliability associated with these content validity assessments was .71 in the first case and .79 in the second.

B29989